UNIVERSIDADE FEDERAL DO PARANÁ

GABRIEL DOS SANTOS MONTEIRO

MINERAÇÃO DE DADOS E MINERAÇÃO DE PROCESSOS EM UMA BASE DE DADOS DE SEGURO GARANTIA

CURITIBA
2017

GABRIEL DOS SANTOS MONTEIRO

MINERAÇÃO DE DADOS E MINERAÇÃO DE PROCESSOS EM BASES DE SEGURO GARANTIA

Trabalho de Conclusão de Curso apresentado como requisito parcial à obtenção do grau de Bacharel em Gestão da Informação no curso de graduação em Gestão da Informação, Setor de Ciências Sociais Aplicadas da Universidade Federal do Paraná.

Orientadora: Profª. Drª. Denise Fukumi Tsunoda

CURITIBA
2017

TERMO DE APROVAÇÃO

GABRIEL DOS SANTOS MONTEIRO

MINERAÇÃO DE DADOS E MINERAÇÃO DE PROCESSOS EM UMA BASE DE DADOS DE SEGURO GARANTIA

Trabalho de Conclusão de Curso apresentado como requisito parcial à obtenção do grau de Bacharel em Gestão da Informação no curso de graduação em Gestão da Informação, Setor de Ciências Sociais Aplicadas da Universidade Federal do Paraná, pela seguinte banca examinadora:

_____

Prof.ª Dr. ª Denise Fukumi Tsunoda
Orientadora - Setor de Ciências Sociais Aplicadas da Universidade Federal, UFPR

_____

Prof. Me. André José Ribeiro Guimarães
Setor de Ciências Sociais Aplicadas da Universidade Federal, UFPR

_____

Prof. Dr. Cícero Aparecido Bezerra
Setor de Ciências Sociais Aplicadas da Universidade Federal, UFPR

Curitiba, 06 de dezembro de 2017

RESUMO

Avalia a identificação de padrões em processos de subscrição de seguro por meio da mineração de dados tradicional e da mineração de processos em uma base de dados extraída do sistema utilizado em uma organização do setor de seguro garantia em Curitiba. Aplica-se o algoritmo PART para classificação do resultado final do processo e apresenta as principais regras geradas pelo algoritmo. Aplica-se também o algoritmo Fuzzy Miner para mineração de processos, gerando diferentes modelos de processos com fluxos e atividades baseados nos parâmetros definidos. Apresenta como principal modelo aquele que apresenta todos os fluxos realizados pelos processos e abstrai as atividades incluindo apenas as mais frequentes. Define como trabalho futuro a validação dos resultados obtidos com usuários do sistema e especialistas da área de seguros.

Palavras-chave: Seguro Garantia. Gestão da Informação. Descoberta de Conhecimento em Bases de Dados. Mineração de Dados. Mineração de Processos.

ABSTRACT

Evaluates the discovery of patterns in insurance underwriting processes through traditional data mining and process mining in a database consisted of processes from an insurance company in Curitiba. PART algorithm is used to classify the final result of the process and presents the main rules generated by it. The Fuzzy Miner algorithm for process mining is also used, generating different process models with flows and activities based on defined parameters. Presents as main model the one that presents all the flows executed by the processes and abstracts the activities including only the most frequent ones. Defines as future research the validation of the obtained results with system users and insurance specialists.

Keywords: Surety Bond. Information Management. Knowledge Discovery In Databases. Data Mining. Process Mining.

LISTA DE FIGURAS

LISTA DE QUADROS

SUMÁRIO

1 INTRODUÇÃO

O século XXI trouxe um grande avanço na tecnologia, principalmente quando se trata de informação. É cada vez mais fácil utilizar e compartilhar informações de qualquer parte do mundo a qualquer hora. Com isso, o volume de informações naturalmente aumentou e viu-se a necessidade de gerir e tratar essas informações de modo a extrair utilidade das mesmas, seja para tomada de decisão, resolução de problemas ou inovações de variados tipos. Essa mudança também tornou obsoleta a maioria dos métodos tradicionais de análise de dados e informações quando se trata de grandes quantidades.

Surge assim a mineração de dados como alternativa para manipular essa massa de dados a fim de auxiliar as análises através de diferentes algoritmos e técnicas que buscam padrões e tendências que podem levar a resultados conclusivos.

Juntamente com a mineração de dados, surge também a mineração de processos. As organizações de hoje têm como suporte para suas atividades uma grande variedade de sistemas de informação. Em muitos casos, esses sistemas geram registros de eventos que podem ser ligados a execução de tarefas dos processos organizacionais. Da mesma forma que a mineração de dados é aplicada a grandes volumes de dados, a mineração de processos é aplicada aos *logs* destes eventos para obter análises das execuções de processos e extrair conhecimento acerca dos mesmos.

Tendo em vista as finalidades aqui descritas da mineração de dados e da mineração de processos, o presente projeto visa a utilização de ambas as abordagens para analisar bases de dados relativas a processos de subscrição de seguro garantia e buscar identificar padrões nas etapas destes processos e seus resultados finais, assim como analisar o modelo de processo seguido em comparação com seu modelo idealizado e possíveis desvios dele. O processo de subscrição consiste na solicitação de apólices de seguro feita por clientes junto à organização, seguido pela análise desta solicitação de acordo com as variáveis pertinentes ao processo prático e legal envolvido (feita por diferentes departamentos, de acordo com as variáveis analisadas) e posteriormente a decisão final de aceitação ou recusa da subscrição de seguro, que pode vir acompanhada ou não de restrições.

1.1 PROBLEMATIZAÇÃO

Uma grande ou média empresa do setor de seguros lida com uma grande quantidade de solicitações de subscrição e emissões de apólices, passando por diversos setores para diferentes tipos de análise até seu resultado final. Cada uma dessas solicitações gera um amplo volume de dados e metadados, tanto das solicitações em si quanto do seu processo de análise. Isso torna complexo o processo de análise de dados e informações, dificultando a visualização de padrões que podem auxiliar em melhorias ou fornecer soluções para problemas.

A partir deste cenário, nota-se que é importante utilizar a recuperação da informação de maneira proveitosa diante deste grande volume de registros gerados. Da mesma forma, faz-se valoroso buscar o estabelecimento de mecanismos automáticos de processamento a fim de tornar o processo de descoberta de conhecimento mais eficiente na exploração de bases de dados, execuções de processos e reconhecimento de padrões existentes nos mesmos.

Sendo assim, define-se a questão de pesquisa a ser respondida por esse projeto: **Como verificar padrões em resultados de processos de subscrição de seguro garantia por meio da aplicação de técnicas de mineração de dados e mineração de processos?**

1.2 OBJETIVOS

Tendo em vista o problema de pesquisa descrito, os objetivos deste projeto foram definidos e divididos em um objetivo geral e quatro objetivos específicos.

1.2.1 Objetivo Geral

Este projeto tem como objetivo geral identificar padrões em processos de subscrição em modalidades de Seguro Garantia através da aplicação de técnicas de mineração de dados e mineração de processos em uma base de dados desta modalidade de seguro.

1.2.2 Objetivos Específicos

Os objetivos específicos definidos a fim de alcançar o objetivo geral são os seguintes:

- definir, dentre os métodos mais citados na literatura científica, quais os mais adequados para a mineração de dados da base em questão;

- definir uma ou mais ferramentas de mineração de processos a ser(em) utilizadas neste projeto;

- preparar a base para descoberta de padrões em conformidade com os métodos e ferramentas escolhidos;

- realizar a análise descritiva da base de dados e analisar os resultados obtidos.

1.3 JUSTIFICATIVA

Esta seção apresenta a justificativa sob três aspectos, a saber: científica (área de pesquisa), acadêmica (para o curso de Gestão da informação) e pessoal.

1.3.1 Para a área de pesquisa

Foi realizado um levantamento em 28 de março de 2017 nas bases *Web of Science* e *Scopus* a fim de verificar o estado atual do campo de pesquisas nesta área considerando o volume do acervo disponível. A Figura 1 mostra os termos utilizados no levantamento e seus respectivos resultados. É importante destacar que em ambas as bases pesquisadas não se obteve nenhum resultado para o cruzamento dos termos “data mining” e “surety bond” assim como para o cruzamento dos termos “process mining” e “surety bond”, o que indica uma grande escassez de pesquisa desenvolvida nesta área de estudo, reforçando a relevância da contribuição do desenvolvimento deste estudo.

FIGURA 1 - NÚMERO DE RESULTADOS DE BUSCAS NAS BASES WEB OF SCIENCE E SCOPUS

Web of Science

Termos	Resultados
"data mining"	11.279
"data mining" & "process"	217
"process mining"	387
"data" & "process mining"	387
"data mining" & "process mining"	1
"surety bond"	5
"data" & "surety bond"	1
"data mining" & "surety bond"	0
"process mining" & "surety bond"	0

Scopus

Termos	Resultados
"data mining"	49.521
"data mining" & "process"	25.728
"process mining"	1.876
"data" & "process mining"	1.432
"data mining" & "process mining"	1.063
"surety bond"	52
"data" & "surety bond"	6
"data mining" & "surety bond"	0
"process mining" & "surety bond"	0

FONTE: WEB OF SCIENCE (2017) E SCOPUS (2017)

A Figura 2 mostra as 15 principais áreas de estudo da base *Web of Science* onde se desenvolveram pesquisas em mineração de dados.

FIGURA 2 - ÁREAS DE PESQUISA PREDOMINANTES PARA O TERMO “DATA MINING”

Campo: Áreas de pesquisa	Contagem do registro	% de 11312	Gráfico de barras
COMPUTER SCIENCE	7003	61.908 %	
ENGINEERING	3459	30.578 %	
OPERATIONS RESEARCH MANAGEMENT SCIENCE	670	5.923 %	
TELECOMMUNICATIONS	551	4.871 %	
AUTOMATION CONTROL SYSTEMS	541	4.783 %	
BUSINESS ECONOMICS	512	4.526 %	
MATHEMATICS	455	4.022 %	
BIOCHEMISTRY MOLECULAR BIOLOGY	286	2.528 %	
MEDICAL INFORMATICS	268	2.369 %	
PHARMACOLOGY PHARMACY	258	2.281 %	
MATHEMATICAL COMPUTATIONAL BIOLOGY	251	2.219 %	
MATERIALS SCIENCE	242	2.139 %	
CHEMISTRY	241	2.130 %	
INFORMATION SCIENCE LIBRARY SCIENCE	222	1.963 %	
EDUCATION EDUCATIONAL RESEARCH	212	1.874 %	

FONTE: WEB OF SCIENCE (2017)

A Figura 3 mostra as 10 principais áreas de estudo da base *Web of Science* onde se desenvolveram pesquisas em mineração de processos.

FIGURA 3 - ÁREAS DE PESQUISA PREDOMINANTES PARA O TERMO “PROCESS MINING”

Campo: Áreas de pesquisa	Contagem do registro	% de 387	Gráfico de barras
COMPUTER SCIENCE	323	83.463 %	
ENGINEERING	94	24.289 %	
OPERATIONS RESEARCH MANAGEMENT SCIENCE	36	9.302 %	
TELECOMMUNICATIONS	24	6.202 %	
BUSINESS ECONOMICS	19	4.910 %	
AUTOMATION CONTROL SYSTEMS	13	3.359 %	
MEDICAL INFORMATICS	12	3.101 %	
HEALTH CARE SCIENCES SERVICES	7	1.809 %	
ROBOTICS	7	1.809 %	
EDUCATION EDUCATIONAL RESEARCH	6	1.550 %	

FONTE: WEB OF SCIENCE (2017)

1.3.2 Para o curso de gestão da informação

Além da falta de estudos desenvolvidos na área como um todo, outra motivação para a realização deste estudo é a falta de pesquisas no curso de Gestão da Informação, curso onde é atualmente desenvolvida a graduação referente a este projeto. O desenvolvimento deste seria de grande importância considerando seu ineditismo e possível incentivo para novas pesquisas dentro do curso.

1.3.3 Para o autor

Por fim, o período de estágio desenvolvido em uma organização atuante no setor foi também determinante para a motivação deste projeto. Ao vivenciar a análise dos processos de subscrição no dia-a-dia, surgiu o interesse pessoal em aplicar técnicas e ferramentas de mineração de dados e de processos que pudessem auxiliar na identificação de padrões para otimizar a tomada de decisão.

1.4 DELIMITAÇÕES DA PESQUISA

A pesquisa abrange uma base de dados cedida por uma grande empresa de Curitiba atuante no setor de seguro garantia. A base em questão traz processos de subscrição em suas várias fases, incluindo principalmente o tempo de duração de cada etapa.

1.5 ESTRUTURA DO DOCUMENTO

O documento está dividido em seis seções principais. A primeira seção apresenta o parecer introdutório sobre o estudo realizado, apresentando brevemente conceitos importantes para a área estudada; o problema de pesquisa que levou à realização deste estudo, assim como a justificativa de sua realização, por fim delimitando o escopo da pesquisa.

A segunda seção apresenta o referencial teórico que embasou este estudo, abrangendo o conceito de recuperação da informação nos seus diferentes tipos, a mineração de dados e suas principais técnicas e respectivos algoritmos, a mineração de processos e principais ferramentas disponíveis para sua realização, e finalmente o conceito de seguro garantia e como esta modalidade de seguros se diferencia das demais.

A terceira seção apresenta a metodologia aplicada na realização deste estudo. Aponta a caracterização da pesquisa, assim como os métodos e ferramentas usados na mineração de dados, mineração de processos e na análise estatística descritiva da base de dados.

A quarta seção apresenta os resultados da mineração de dados tradicional, assim como da mineração de processos. Apresenta também a análise estatística descritiva da base de dados.

Por fim, a quinta seção apresenta as considerações finais do projeto, o alcance dos objetivos definidos e trabalhos futuros.

2 REFERENCIAL TEÓRICO

Esta seção apresenta a revisão da literatura que fundamenta a pesquisa e proposta deste estudo de acordo com a problemática investigada e os objetivos traçados. Serão abordados como temas: recuperação da informação, KDD (Knowledge Discovery in Databases), mineração de dados, mineração de processos, seguro garantia e a relação entre as áreas de mineração de dados e de seguro garantia.

2.1 RECUPERAÇÃO DA INFORMAÇÃO

Baeza-Yates e Ribeiro-Neto (2013) tratam a Recuperação da Informação como uma área abrangente que se concentra principalmente em prover aos usuários o acesso fácil às informações de seu interesse. A definição dada pelos autores é a seguinte:

> A Recuperação de Informação (RI) trata da recuperação, armazenamento, organização e acesso a itens de informação, como documentos, páginas Web, catálogos online, registros estruturados e semiestruturados, objetos multimídia, etc. A representação e a organização dos itens devem fornecer aos usuários facilidade de acesso às informações de seu interesse. (BAEZA-YATES; RIBEIRO-NETO, 2013, p. 1)

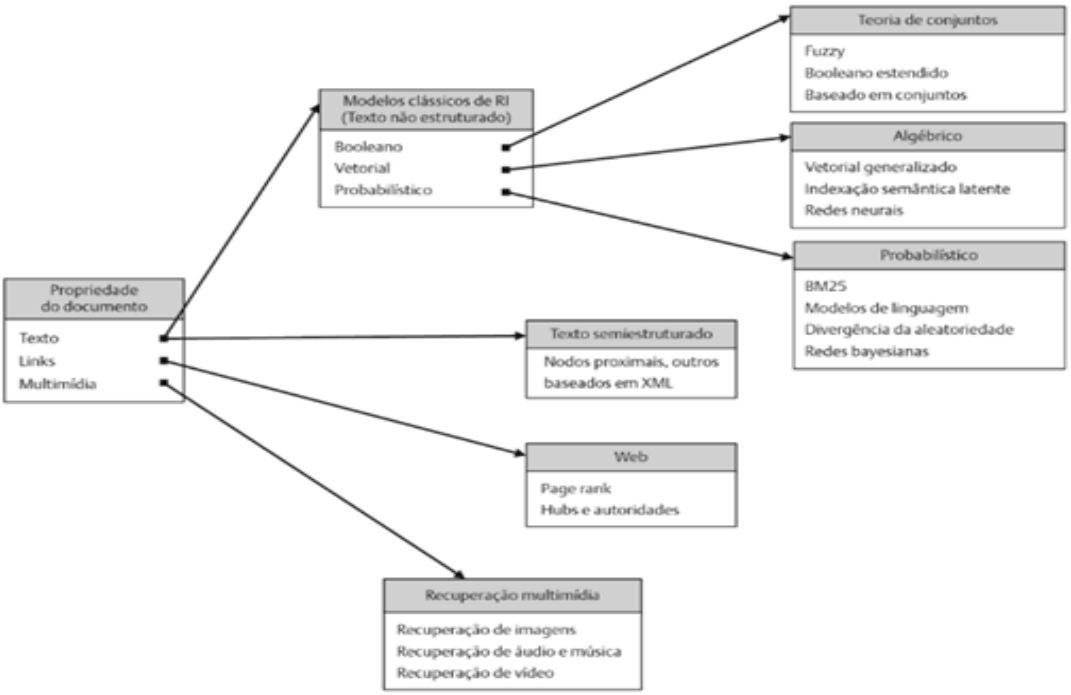
Ainda de acordo com Baeza-Yates e Ribeiro-Neto (2013), em termos de pesquisa, a RI pode ser estudada sob dois pontos de vista diferentes, porém que se adicionam: um deles é centrado no computador é o outro é centrado no usuário. Os autores afirmam que na visão centrada no computador, a RI consiste principalmente na construção de índices eficientes, no processamento de consultas com alto desempenho e no desenvolvimento de algoritmos de ranqueamento, a fim de melhorar os resultados. Já na visão centrada no usuário, afirmam que a RI foca no estudo do comportamento do usuário, buscando entender suas principais necessidades e determinar como esse entendimento afetar a organização e a operação do sistema de recuperação. Afirmam também que a visão centrada no computador é dominante no meio acadêmico e no mercado.

Ainda de acordo com os autores, os principais modelos de RI funcionam atribuindo pontuações a documentos em relação a uma consulta e faz um ranqueamento através desta pontuação. Esses modelos são fundamentalmente

baseados em texto, isto é, eles usam o texto dos documentos para ranqueá-los em relação à consulta. Na Web, contudo, os autores destacam que também é necessário utilizar a informação sobre a estrutura de links para alcançar um bom rastreamento. Porém, quando se trata de objetos multimídia, a codificação é feita de de forma diferente. Imagens são codificadas como *bitmaps* de *pixels*, vídeos são codificados como fluxos (*streams*) temporais de imagens e objetos de áudio são codificados como fluxos discretizados de som. Por possuírem essas diferentes formas de representação, os objetos multimídia são ranqueados de maneira diferente, ou então são recuperados sem ranqueamentos. Dadas essas características, os autores distinguem três categorias de modelos de RI: baseadas em texto, as baseadas em links e as baseadas em objetos de multimídia.

A Figura 4 ilustra a taxonomia de modelos de RI elaborada pelos autores, bem como os modelos de recuperação de objetos de multimídia. Quanto aos modelos baseados em texto, os autores distinguem entre modelos para texto não estruturado e modelo que levam em conta a estruturação do texto.

FIGURA 4 - UMA TAXONOMIA DOS MODELOS DE RI



FONTE: BAEZA-YATES E RIBEIRO-NETO (2013, P. 24)

Na primeira categoria do modelo, o texto é modelado como uma simples

sequência de palavras. Na segunda categoria, componentes estruturais do texto (como título, seções, subseções e parágrafos) são uma parte integral do modelo, que geralmente é chamado de semiestruturado, porque inclui ambos textos estruturados e não estruturados. Quanto ao texto não estruturado, os três modelos clássicos são chamados de Booleano, vetorial e probabilístico. No modelo Booleano, documentos e consultas são representados como conjuntos de termos de indexação, sendo denominado pelos autores como modelo da teoria de conjuntos. No modelo vetorial, documentos e consultas são representados come vetores em um espaço com t dimensões, sendo definido pelos autores como o modelo algébrico. No modelo probabilístico, o arcabouço para modelar as representações dos documentos e consultas é baseado na teoria das probabilidades. Dessa forma, os autores chamam este modelo de probabilístico. (BAEZA-YATES; RIBEIRO-NETO, 2013)

Baeza-Yates e Ribeiro-Neto (2013) também apresentam outros modelos alternativos. Quanto aos modelos alternativos baseados na teoria dos conjuntos, os autores apresentam o *fuzzy,* o booleano estendido e o baseado em conjuntos. Quanto aos modelos algébricos alternativos, trazem o modelo vetorial generalizado, a indexação semântica latente e o modelo de redes neurais. Quanto aos modelos probabilísticos alternativos, distinguem o BM25, o de redes bayesianas, a divergência da aleatoriedade e os modelos de linguagem.

Em relação aos modelos para a recuperação de textos semiestruturados (isto é, modelos que lidam com a estrutura fornecida pelo texto), os autores consideram técnicas de indexação como os nodos proximais e os métodos de indexação baseados em XML. Já para a Web, devido ao grande número de documentos (ou páginas Web), o ranqueamento baseado em texto por si só não é suficiente. Também é necessário considerar os *links* entre páginas Web como parte integrante do modelo.

Os autores afirmam que um conjunto diferente de estratégias de recuperação é empregado para dados multimídia. Para recuperar imagens de interesse do usuário, são necessários vários passos intermediários que não são requeridos na busca em coleções textuais. Os autores exemplificam que, em vez de escrever uma consulta, o usuário pode especificar sua necessidade de informação apontando para uma dada imagem. Essa imagem consultada é comparada pelos autores às imagens da coleção para recuperar imagens relacionadas. Assim, os autores explicam que os métodos para a recuperação multimídia são muito distintos dos métodos de RI para texto, pois,

por exemplo, muitos deles não incluem nenhuma forma de ranqueamento. Por essa razão, os autores afirmam que eles são representados separadamente pelos autores na taxonomia.

Por fim, Baeza-Yates e Ribeiro-Neto (2013) consideram que o propósito principal de um modelo de RI é produzir um conjunto de resultados que provavelmente seja relevante para o usuário, implementações modernas de sistemas de RI incluem características de vários modelos de RI, e não de apenas um. Por exemplo, funções de ranqueamento na Web combinam características dos modelos clássicos de RI com características de modelos baseados em links para melhorar a recuperação.

A recuperação de informação também pode ser realizada através de técnicas de mineração de dados. A próxima subseção apresenta essa abordagem.

2.2 MINERAÇÃO DE DADOS

A abordagem clássica de análise de dados geralmente consiste em um ou mais analistas explorando manualmente um conjunto de dados. Porém este processo manual é lento, caro, altamente subjetivo e pode apresentar mais falhas quando comparado ao processo automatizado.

Por isso, visando otimizar a obtenção de conhecimento através de dados, utiliza-se o processo chamado de Descoberta de Conhecimento em Bancos de Dados (Knowledge Discovery in Databases - KDD) do qual é parte importante a etapa de Mineração de Dados.

Esta etapa do KDD consiste em aplicar algoritmos de análise de dados e de descobertas com o objetivo de extrair conhecimento implícito por meio da descoberta de padrões e regras significativas, a partir de grande quantidade de dados armazenados, de forma automática ou semiautomática, utilizando modelos computacionais construídos para descobrir novos fatos e relacionamentos entre dados, de forma repetida e interativa (FAYYAD et al., 1996, p. 1).

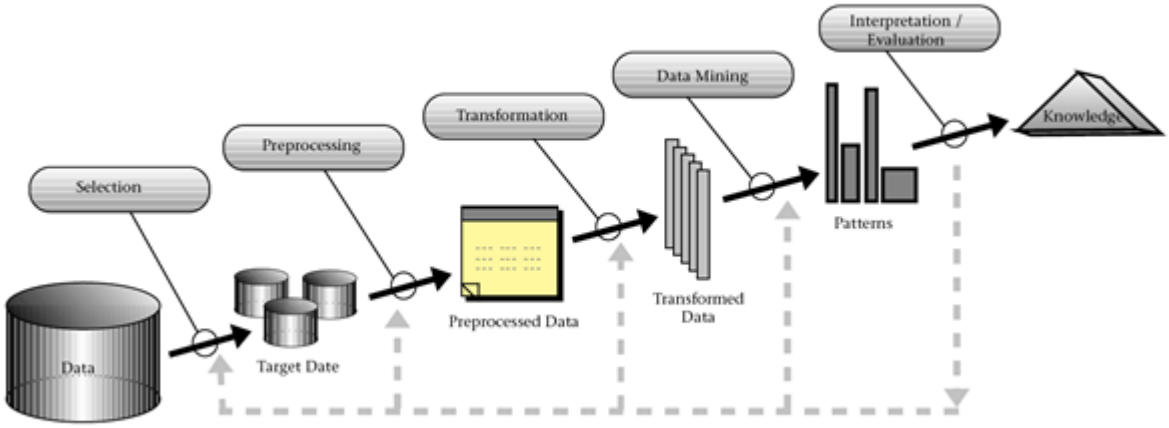
As etapas do processo KDD como um todo serão descritas no próximo tópico.

2.2.1 Knowledge Discovery In Databases (KDD)

De acordo com Fayyad et al. (1996), KDD é “o processo, não trivial, de extração de informações implícitas, previamente desconhecidas e potencialmente úteis, a partir dos dados armazenados em um banco de dados”. Este processo é necessário devido ao grande avanço da tecnologia e consequentemente maior volume e utilização de dados, tornando mais complexo o entendimento de tais dados sem que haja um filtro do que é realmente útil.

Fayyad et al. (1996) também define o KDD em cinco fases: seleção; pré-processamento; transformação; mineração de dados; e interpretação/avaliação. Estas fases estão representadas na Figura 5.

FIGURA 5 - ETAPAS DO KDD



FONTE: FAYYAD ET AL. (1996, P. 41)

Estas etapas do processo de KDD podem ser descritas como:

- seleção de dados: é escolhido o conjunto de dados contendo todas as possíveis variáveis (atributos) e registros que farão parte da análise;
- pré-processamento: aqui, faz-se o tratamento dos dados previamente selecionados, realizando a limpeza dos dados e a remoção de ruídos para garantir a qualidade dos dados e a eficiência dos algoritmos aplicados;
- transformação: nesta etapa, transforma-se os dados de acordo com a objetivo da aplicação da mineração. é também nesta etapa que ocorrem os

processos de armazenamento e formatação para que os algoritmos possam ser aplicados adequadamente;

- mineração de dados: são aplicados os algoritmos escolhidos de acordo com as tarefas de mineração apropriadas, assim como a parametrização dos mesmos tendo em visto o conjunto de dados-alvo;
- interpretação e avaliação dos resultados: por fim, os resultados são interpretados e avaliados.

Nota-se que uma das etapas mais importantes do processo KDD é a mineração de dados. Nesta etapa, existem diferentes tarefas realizadas por algoritmos de acordo com os objetivos da aplicação. As tarefas mais importantes de algoritmos de extração de padrões serão descritas na seção a seguir.

2.2.2 Tarefas e métodos de mineração de dados

Fayyad et al. (1996) diz que os dois objetivos mais notáveis da mineração de dados tendem a ser a predição e a descrição. Esses dois objetivos podem ser atingidos usando uma variedade de métodos de mineração de dados, divididos aqui em algoritmos de classificação, associação, agrupamento e regressão.

2.2.2.1 Classificação

As técnicas de classificação são definidas por Fayyad et al. (1996) como técnicas que realizam o mapeamento de um registro de dados de acordo em uma de diferentes classes pré-definidas. Entre os principais algoritmos de classificação presentes na literatura, destacam-se o ID3, o C4.5, o PRISM, as tabelas de decisão, as redes neurais e os algoritmos genéticos, descritos a seguir.

O algoritmo ID3 é um algoritmo de classificação através de árvore de decisão. Amo (s/d) diz que uma árvore de decisão é uma estrutura de árvore onde:

- cada nó interno é um atributo do banco de dados de amostras, diferente do atributo-classe;
- as folhas são valores do atributo-classe;
- cada ramo ligando um nó-filho a um nó-pai é etiquetado com um valor do

atributo contido no nó-pai. Existem tantos ramos quantos valores possíveis para este atributo;

- um atributo que aparece num nó não pode aparecer em seus nós descendentes.

Camilo e Silva (2009) afirmam que “o sucesso das árvores de decisão, deve-se ao fato de ser uma técnica extremamente simples, não necessita de parâmetros de configuração e geralmente tem um bom grau de assertividade”.

O algoritmo C4.5 é um dos sucessores do ID3. Assim como o ID3, o C4.5 gera classificadores expressados em árvores de decisão, mas também pode gerá-los de forma de regras mais compreensíveis. Além disso, as operações de poda suportadas pelo C4.5 geralmente resultam em classificadores que não podem ser reformulados árvores de decisão (WU; KUMAR, 2009).

O algoritmo PRISM, embora também baseado no ID3, utiliza uma diferente estratégia de indução, o qual gera regras modulares, consideradas independentes e de fácil compreensão. Este algoritmo possui em sua entrada um conjunto de atributos e gera como resultado um conjunto de regras de classificação no formato SE > ENTÃO (CENDROWSKA, 1987).

De acordo com Camilo e Silva (2009), as redes neurais têm origem na psicologia e na neurobiologia. Afirmam ainda que esta técnica consiste basicamente em simular o comportamento dos neurônios. Uma rede neural pode ser vista como um conjunto de unidades de entrada e saída conectados por camadas intermediárias e uma dessas conexões possui um peso associado. Durante o processo de aprendizado, a rede ajusta estes pesos para conseguir classificar corretamente um objeto.

Este algoritmo de classificação terá como entrada um banco de dados de treinamento e retornará como resultado uma rede neural. Esta rede também poderá ser transformada num conjunto de regras de classificação, como acontece com as árvores de decisão (AMO, s/d).

O algoritmo de tabelas de decisão (*decision tables*), assim como as árvores de decisão e as redes neurais, gera modelos de classificação utilizados para previsão. É induzida por algoritmos de *machine learning*. Uma *decision table* é tabela com uma estrutura hierárquica em que cada registro no nível mais alto da tabela dá origem a

uma outra tabela menor, com base na correspondência dos valores de diferentes atributos (KOHAVI, 1995).

Os algoritmos genéticos funcionam como uma técnica que segue a teoria da evolução. Geralmente, define-se no estágio inicial uma população de maneira aleatória. Seguindo a lei do mais forte (assim como na evolução), é gerada uma nova população decorrente da atual, porém, são realizados processos de troca genética e mutação para que os novos indivíduos surjam. Este processo continua até que populações com indivíduos mais fortes sejam geradas ou que atinja algum critério de parada (CAMILO E SILVA, 2009, p. 15).

2.2.2.2 Associação

A tarefa de associação consiste basicamente em identificar quais atributos estão relacionados entre si. Foca em encontrar padrões de fácil interpretação que descrevam os dados. (FAYYAD et al., 1996, p. 44).

O algoritmo Apriori é método mais conhecido para a mineração de regras de associação e emprega busca de profundidade e gera conjuntos de itens candidatos de *n* elementos a partir de um conjunto de itens com *n - 1* elementos (CASTRO; FERRARI, 2016).

2.2.2.3 Regressão

A estimação é similar à classificação, porém é usada quando o registro é identificado por um valor numérico e não um categórico. Assim, pode-se estimar o valor de uma determinada variável analisando-se os valores das demais.

Dentro da regressão podemos ter - entre outras - a linear e a não linear. As regressões são chamadas de lineares quando a relação entre as variáveis preditoras e a resposta segue um comportamento linear. Neste caso, é possível criar um modelo no qual o valor de y é uma função linear de x. Nos modelos de regressão não-linear, a relação entre as variáveis preditoras e a resposta não segue um comportamento linear. Por exemplo, a relação entre as variáveis pode ser modelada como uma função polinomial. (CAMILO E SILVA, 2009, p. 16).

2.2.2.4 Agrupamento

O agrupamento ou *clustering* é uma tarefa descritiva comum, onde o objetivo é especificar um conjunto finito de categorias ou grupos para descrever o conjunto de dados (FAYYAD et al., 1996, p. 44).

O principal algoritmo para esta tarefa é o *k-means*. De acordo com Camilo e Silva (2009), esse algoritmo usa o conceito da centróide. Dado um conjunto de dados, o algoritmo seleciona de forma aleatória k registros, cada um representando um agrupamento. Para cada registro restante, é calculada a similaridade entre o registro analisado e o centro de cada agrupamento. O objeto é inserido no agrupamento com a menor distância, ou seja, maior similaridade. O centro do cluster é recalculado a cada novo elemento inserido.

Uma variação da mineração de dados é a mineração de processos, abordada no tópico seguinte.

2.3 MINERAÇÃO DE PROCESSOS

De acordo com Aaslt (2016), a mineração de processos é uma disciplina de pesquisa relativamente jovem que traz o *machine learning* e a mineração de dados em um lado e a análise e modelagem de processos no outro.

> A mineração de processos pode ser vista como uma maneira de preencher a lacuna entre a ciência de dados e a ciência de processos. A abordagem de ciência de dados tende a deixar de lado os processos enquanto a abordagem da ciência de processos tende a ser focada nos modelos de processos sem considerar a evidência escondida nos dados. (AASLT, 2016, p. 17)

A ideia da mineração de processos é descobrir, monitorar e melhorar processos reais (i.e., não os processos idealizados) ao extrair conhecimentos de *logs* de eventos que muitas vezes já vêm prontos em sistemas atuais. O Quadro 1 mostra um exemplo simples de um *log* de eventos, onde cada evento é identificado com um ID e as tarefas de cada processo aparecem agregadas.

QUADRO 1 - EXEMPLO DE LOG DE EVENTOS

ID	Processo
1	Começar, Fazer matrícula para CNH, Frequentar aulas para carros, Fazer prova teórica, Fazer exame prático para carros, Receber resultado, Finalizar
2	Começar, Fazer matrícula para CNH, Frequentar aulas para motos, Fazer prova teórica, Fazer exame prático para motos, Receber resultado, Receber habilitação, Finalizar
3	Começar, Fazer matrícula para CNH, Frequentar aulas para carros, Fazer prova teórica, Fazer exame prático para carros, Receber resultado, Receber habilitação, Finalizar
4	Começar, Fazer matrícula para CNH, Frequentar aulas para motos, Fazer prova teórica, Fazer exame prático para motos, Receber resultado, Finalizar

FONTE: O AUTOR (2017).

A mineração de processos e suas técnicas permitem a extração de informações de processos de negócio nos *logs* de eventos em sistemas de informação. Esses *logs* geralmente incluem as atividades do processo, responsáveis pela execução das atividades e a data na qual as atividades foram realizadas. Como já dito, um dos objetivos principais da mineração de processos é extrair o modelo real do processo (consistente com o comportamento dinâmico observado no log de eventos) e descobrir como o processo está estruturado e como os atributos de dados influenciam nos pontos de decisão do fluxo de trabalho do processo. Através do modelo gerado é possível também observar quais atividades desencadearam um certo evento ou conjunto de eventos.

FIGURA 6 - VISÃO GERAL DA MINERAÇÃO DE PROCESSOS

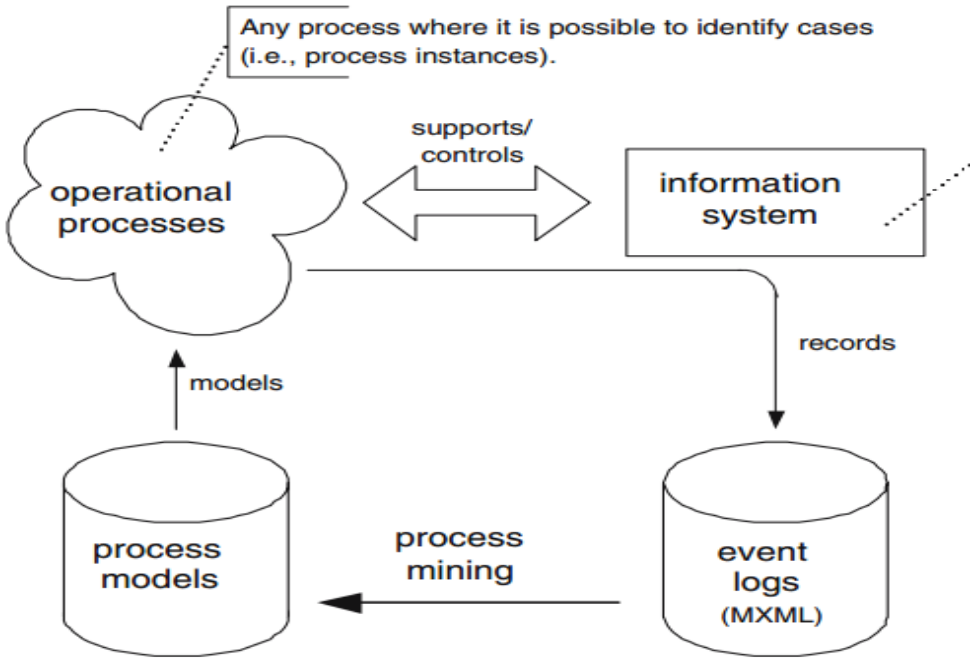


Fig. 1 Overview of process mining

FONTE: MEDEIROS; WEIJTERS; VAN DER AALST (2007)

A Figura 6 mostra uma visão geral de mineração de processos. Processos operacionais (*operational processes*) são suportados ou controlados por um sistema de informação (*information system*) que registra eventos em um *log* de eventos. Esse *log* é utilizado para extrair modelos de processos que descrevem o comportamento observado. Essa informação é valiosa para melhor entender os processos e melhorá-los. Processos reais tendem a desviar do modelo ideal com que foram implementados, por isso a relevância prática da mineração de processos - ao explorar o verdadeiro modelo seguido pelo processo independente daquele idealizado (MEDEIROS; WEIJTERS; van der AALST, 2007).

Aaslt (2016) apresenta três tipos diferentes de mineração de processos que podem ser realizados:

- o primeiro tipo é chamado de *discovery* (descoberta). Ele recebe esse nome pois gera um modelo de processo real a partir de um *log* de eventos sem nenhuma informação de antemão. Recebendo exemplos suficientes de execução do processo, é possível construir automaticamente o modelo sem nenhum conhecimento adicional;

- o segundo tipo de mineração de processos é chamado de *conformance* (conformidade). Neste, um modelo de processo existente é comparado com o *log* de eventos do mesmo processo. A verificação de conformidade pode ser utilizada para verificar se a realidade, conforme o *log* de eventos, está de acordo com o modelo e vice-versa. Por exemplo, imagina-se um modelo de processo que indica as compras abaixo de R$ 1.000,00 em uma loja não podem ser parceladas. A análise do *log* mostrará se essa regra está sendo seguida ou não;
- o terceiro tipo de mineração de processos se chama *enhancement* (melhoria). A ideia principal deste tipo é ampliar ou melhorar um modelo de processo existente utilizando informações acerca do processo real registrado em um *log* de eventos. Enquanto a verificação de conformidade mede o alinhamento entre o modelo e a realidade, este terceiro tipo de mineração de processos foca na mudança ou melhoria do modelo já existente. Um tipo de melhoria que pode ser feita é o reparo, que modifica o modelo para refletir melhor a realidade. Outro tipo de melhoria é a extensão, que adiciona uma nova perspectiva ao modelo do processo a partir do cruzamento de informações com o *log* de eventos.

Ainda de acordo com van der Aalst (2016), um dos elementos chave da mineração de processos é a ênfase em estabelecer uma forte relação entre um modelo de processo e a realidade capturada na forma de um *log* de eventos. Essa relação pode ser refletida de três formas diferentes, chamadas de *Play-Out*, *Play-In* e Replay.

A abordagem *Play-Out* se refere ao uso clássico dos modelos de processos. A partir de um modelo, é possível gerar seu comportamento esperado. Essa abordagem pode ser utilizada para simular modelos e conduzir experimentos acerca daquele modelo específico. A ideia principal da simulação é coletar dados que permitam analisar o que se pode esperar ao colocar o modelo em prática.

A abordagem *Play-In* é o oposto da abordagem *Play-Out*, é tomado o comportamento de um processo como entrada e o objetivo é construir um modelo. Esta abordagem pode também ser chamada de inferência e um exemplo é apresentado abaixo. O *log* de eventos na Tabela 1 dá origem ao modelo de processo

inferido mostrado na Figura 7.

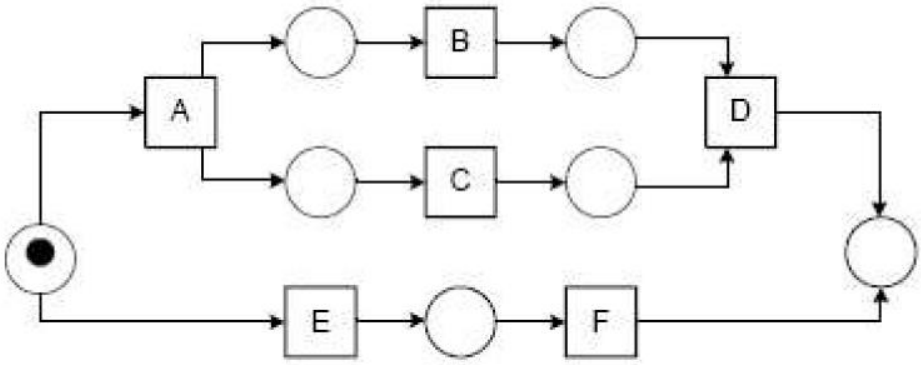
QUADRO 2 - EXEMPLO DE LOG DE PROCESSO PARA INFERÊNCIA

ID Caso	ID Tarefa
Caso 1	Tarefa A
Caso 2	Tarefa A
Caso 3	Tarefa A
Caso 3	Tarefa B
Caso 1	Tarefa B
Caso 1	Tarefa C
Caso 2	Tarefa C
Caso 4	Tarefa A
Caso 2	Tarefa B
Caso 2	Tarefa D
Caso 5	Tarefa E
Caso 4	Tarefa C
Caso 1	Tarefa D
Caso 3	Tarefa C
Caso 3	Tarefa D
Caso 4	Tarefa B
Caso 5	Tarefa F
Caso 4	Tarefa D

FONTE: van der Aalst; Weijters (2004)

A maioria das técnicas de mineração de dados utiliza uma abordagem parecida: um modelo é gerado com base em exemplos ou conjuntos de dados de treinamento. Infelizmente, não é possível empregar técnicas de mineração de dados tradicionais em modelos de processos com a abordagem *Play-In* devido às suas limitações em trabalhar com *logs* de eventos para esse propósito.

FIGURA 7 - MODELO DE PROCESSO GERADO A PARTIR DO LOG DE EVENTOS DA TABELA 1



FONTE: VAN DER AALST; WEIJTERS (2004)

A abordagem *Replay* utiliza tanto *logs* de eventos quanto modelos de processos como entrada, fazendo com que o *log* seja replicado em cima do modelo de processo. Essa réplica é feita simulando a ordem das tarefas executadas de acordo com os registros *log*. Van der Aalst (2016) diz que este processo pode ser feito para três diferentes propósitos:

- verificação de conformidade: discrepâncias entre o *log* e o modelo podem ser detectadas e quantificadas ao simular o *log*. Por exemplo, a execução do *log* pode mostrar que uma tarefa A e uma tarefa C foram executadas, mas a tarefa B que deveria ter sido executada entre elas não foi.
- extensão do modelo do processo através de frequências e informações temporais: ao simular o *log*, é possível verificar quais partes do modelo são executadas frequentemente ou onde são encontrados gargalos nos processos.
- construção de modelos de previsão: ao simular *logs* de eventos, é possível construir modelos de previsão que antecipam certos casos de acordo com cada tarefa do processo. Por exemplo, um modelo pode prever através de um *log* que sempre que uma certa tarefa X é executada, leva no mínimo 24 horas até seu fim.

2.4 SEGURO GARANTIA

O seguro garantia é uma modalidade de seguros que é pouco conhecida pela

maioria das empresas e consumidores por se tratar de um produto da área de seguros que tem uma aplicação específica não necessária em muitos dos setores da economia. Este produto é destinado a órgãos públicos da administração direta e indireta (federais, estaduais e municipais) que por força de norma legal devem exigir garantias de manutenção de oferta (em caso de concorrência) e de fiel cumprimento dos contratos (FUNENSEG, 2001; POLETTO, 2004). O Seguro Garantia é também aplicável a empresas privadas que, nas suas relações contratuais com terceiros (fornecedores, prestadores de serviços e empreiteiros de obras), desejam anular o risco de descumprimento (desempenho). O custo para o empreendedor é significativamente menor. A manutenção da garantia é feita por meio de pagamento de prêmios (mensais ou anuais) reduzindo significativamente os custos de oportunidade.

Esta modalidade de seguro começou a surgir no Brasil a partir de 1964 com o grande desenvolvimento econômico que vinha acontecendo naquela época. Esse desenvolvimento levou o Estado Brasileiro a execução de grandes obras através de contratos públicos e necessitava garantir a realização destas obras através de sistemas de garantias, então, o Seguro Garantia aparece como uma opção para esta garantia. Porém, o mesmo ainda não era um ramo aprovado no Brasil, fazendo com que fosse necessário buscar a solução contratando esta modalidade de seguro em uma companhia Sueca (FUNENSEG, 2001, p. 106).

Apesar deste interesse pela modalidade de seguro, a mesma só veio a ser regularizada em 23 de maio de 1997, com circular da SUSEP (Superintendência de Seguros Privados) N° 005, de 23 de maio de 1997, onde foram aprovados os modelos de apólice Seguro Garantia, as condições da garantia, as disposições tarifárias, modalidades, taxas básicas e condições contratuais gerais para a emissão de apólice.

3 ENCAMINHAMENTOS METODOLÓGICOS

Nesta seção, serão apresentados o ambiente de desenvolvimento desta pesquisa, sua caracterização, a descrição da base de dados trabalhada, assim como os métodos e ferramentas a ser utilizados para a realização de tal.

3.1 CARACTERIZAÇÃO DA PESQUISA

Do ponto de vista de sua natureza, esta pesquisa pode ser classificada como aplicada, pois segundo Silva (2005), pesquisa aplicada é aquela que objetiva gerar conhecimentos para aplicação prática e dirigidos à solução de problemas específicos, envolvendo verdades e interesses locais ao invés de universais.

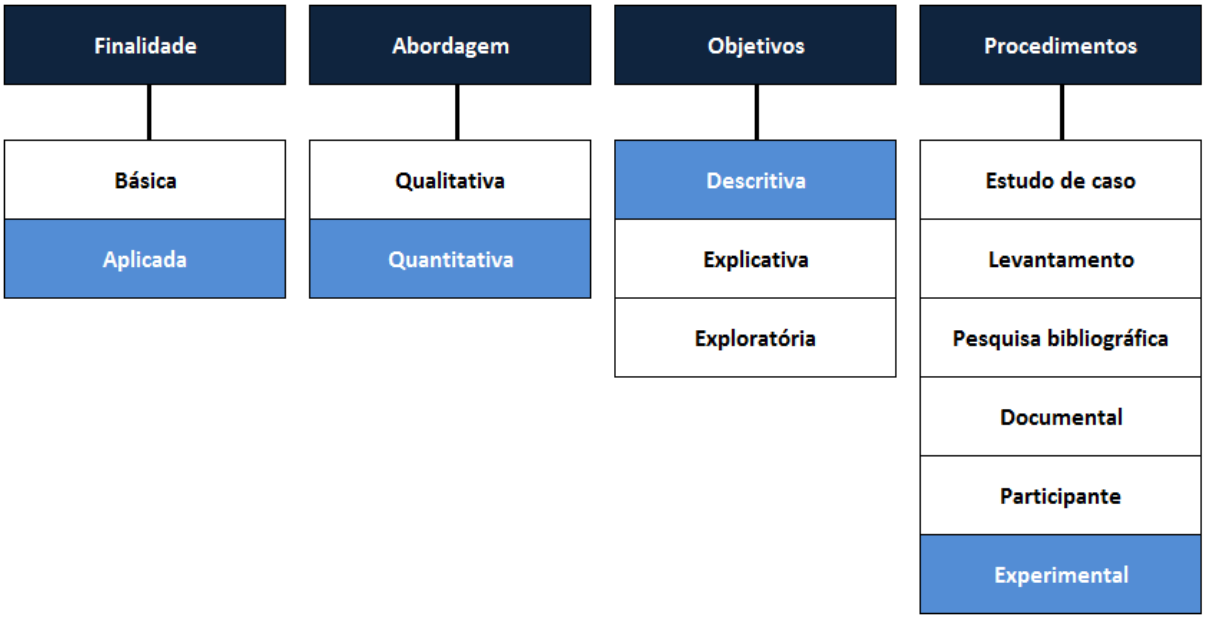
Quanto ao ponto de vista da forma de abordagem do problema, uma pesquisa pode ser caracterizada como qualitativa ou quantitativa. Considerando os aspectos dessa pesquisa, ela pode ser classificada como quantitativa, que é definida por Silva (2005) como uma pesquisa que considera que tudo pode ser quantificável, o que significa traduzir em números opiniões e informações para classificá-las e analisá-las.

Em relação a seus objetivos, Silva (2005) lista três tipos de pesquisa: exploratória, descritiva ou explicativa. A partir destes tipos, é possível afirmar que a presente pesquisa é descritiva, pois “visa descrever as características de determinada população ou fenômeno ou o estabelecimento de relações entre variáveis”. Complementa que este tipo de pesquisa busca descrever um fenômeno ou situação em detalhe e abrange com exatidão as características de um indivíduo, uma situação, ou um grupo, bem como desvenda a relação entre os eventos.

Quanto aos procedimentos técnicos utilizados, definiu-se a pesquisa como experimental, que - de acordo com Silva (2005) - acontece quando se determina um objeto de estudo, selecionam-se as variáveis que seriam capazes de influenciá-lo, definem-se as formas de controle e de observação dos efeitos que a variável produz no objeto.

A Figura 8 mostra a caracterização final da pesquisa de acordo com as classificações abordadas por Silva (2005).

FIGURA 8 - CARACTERIZAÇÃO DA PESQUISA



FONTE: O AUTOR (2017)

As próximas seções detalham os encaminhamentos da pesquisa experimental.

3.2 AMBIENTE DA PESQUISA

A pesquisa será realizada em uma organização privada de Curitiba atuante no setor de seguro garantia.

A organização atua nas diferentes modalidades de seguro garantia se fazendo presente em vários estados do Brasil. Sendo a sua sede em Curitiba, é onde se encontram os dados de todos os processos de subscrição solicitados, seja na sede ou em suas filiais. Para o gerenciamento destes processos, a organização utiliza um sistema próprio que permite realizar o acompanhamento de prazos, verificação das diversas etapas de análises, volume de processos, entre outros.

As bases de dados são extraídas deste sistema quando necessário para análises adicionais utilizando outros softwares com este fim.

Para o desenvolvimento da pesquisa a organização cedeu a sua base de dados de controle do processo, considerando que nenhuma informação pessoal ou ainda que identifique alguma empresa/organização é mencionada nesta pesquisa.

3.3 MATERIAIS E MÉTODOS

Esta seção apresenta a base de dados utilizada neste trabalho, seus atributos, formatos, variações e parte do tratamento realizado antes da execução dos algoritmos de mineração. Apresenta também as ferramentas utilizadas na execução da pesquisa.

3.3.1 Base de dados

A base aqui trabalhada consiste no conjunto de todos os processos e suas etapas já concluídas que tratam de subscrição de seguro de abril de 2014 a junho de 2017, totalizando 27.770 processos. Cada registro da base corresponde a uma etapa de determinado processo; o processo a qual a etapa pertence é determinado pela coluna “Código do CI” presente na base.

A base original retirada da ferramenta Tableau, que é a ferramenta utilizada para gestão dos dados da organização é composta por 34 atributos e 214.992 registros. Os atributos originais da base são:

QUADRO 3 - DESCRIÇÃO DOS ATRIBUTOS ORIGINAIS DA BASE COM SEUS TIPOS E VARIAÇÕES

Atributo	**Tipo de Atributo**	**Valores do Atributo**	**Descrição**
Código do CI	Código numérico	Valores numéricos de até seis algarismos. Ex: 123456	Código que identifica o processo
Soma Tempo Líquido	Tempo no formato hh:mm:ss	Valores de tempo a partir de 00:00:00	Quantia de tempo que durou o processo como um todo
Filtro Cíclico	String de caracteres	Valores categóricos em texto do usuário do processo. Ex: luizpl	Coluna utilizada como filtro interno na ferramenta Tableau
Tempo por Etapa	Tempo no formato hh:mm:ss	Valores de tempo a partir de 00:00:00	Quantia de tempo que durou cada etapa do processo

Etapa	String de caracteres	Valor categórico em texto de uma das 23 etapas possíveis. Ex: 21 - Apólice Emitida.	Descrição da etapa do processo
Status	String de caracteres	Nessa base, assume sempre o valor "Concluído"	Status em que se encontra o processo no sistema
Tempo Líquido	Tempo no formato hh:mm:ss	Valores de tempo a partir de 00:00:00	Medição do tempo apenas para o setor pertinente a esta análise (setor de Subscrição)
Data atualização	Data no formato dd/mm/aaaa	Nessa base, assume sempre o valor 30/01/2015	Data em que a base foi extraída da ferramenta Tableau
% Atingimento	Número em formato decimal	Valores numéricos em formato decimal entre 0 e 1.	Porcentagem de atingimento da meta de tempo do setor
Área	String de caracteres	Valor categórico de uma das duas áreas existentes, Estruturado/Standard ou Tradicionais	Descrição da área do processo
Segundos 18 horas	Número inteiro	Valores numéricos inteiros. Ex: 49800	Contabiliza o número de segundos contidos em dezoito horas
Segundos Inicial	Número inteiro	Valores numéricos inteiros. Ex: 49800	Hora de início do processo convertida em segundos
Segundos Total	Número inteiro	Valores numéricos inteiros. Ex: 49800	Tempo de duração do processo como um todo convertido em segundos
Proposta	Código numérico	Valores numérico de até seis algarismos. Ex: 123456	Proposta de apólice de seguro daquele processo

Segundos 9,5 horas	Número inteiro	Valores numéricos inteiros. Ex: 49800	Contabiliza o número de segundos contidos em nove horas e meia
Segundos 8,5 horas	Número inteiro	Valores numéricos inteiros. Ex: 49800	Contabiliza o número de segundos contidos em oito horas e meia
Segundos Final	Número inteiro	Valores numéricos inteiros. Ex: 49800	Hora de finalização do processo convertida em segundos
Emissor	String de caracteres	Valores categóricos em texto do usuário que emitiu o processo. Ex: luizpl	Usuário que emitiu a apólice do processo
Proposta_Doc	Código numérico	Valores numérico de zero até sete algarismos. Ex: 1234567; 12345; 0	Número da proposta de apólice de seguro contida na apólice de seguro
Tomador_Doc	String de caracteres	Valores de texto com a razão social do Tomador do seguro. Ex: SANESC SANEAMENTO E CONSTRUÇÕES LTDA	Nome do tomador contido na apólice de seguro
Modalidade	String de caracteres	Valores categóricos em texto com uma das 45 modalidades em que o seguro pode acontecer. Ex: Fiança Locatícia	Modalidade de seguro pertinente àquele processo
Inclusor	String de caracteres	Valores categóricos em texto com o usuário que incluiu o processo no sistema Ex: luizpl	Usuário que incluiu o processo no sistema

Conclusor	String de caracteres	Valores categóricos em texto com o usuário que concluiu o processo no sistema. Ex: luizpl	Usuário que concluiu o processo no sistema
Data Final	Data no formato dd/mm/aaaa	Valores de data entre 08/04/2014 e 07/06/2017.	Data final da etapa do processo
Data Inicial	Data no formato dd/mm/aaaa	Valores de data entre 08/04/2014 e 07/06/2017.	Data inicial da etapa do processo
Tomador	String de caracteres	Valores de texto com a razão social do Tomador do seguro. Ex: SANESC SANEAMENTO E CONSTRUÇÕES LTDA	Nome do tomador incluído nas informações iniciais do processo
Passo Encerramento	String de caracteres	Valores categóricos em texto com o passo que encerrou o processo no sistema. Ex: 23 – Oportunidade Perdida	Passo que encerrou o processo
Passo Inicial Etapa	String de caracteres	Valores categóricos em texto com o passo que inicou aquela etapa do processo no sistema. Ex: 3 - Análise do risco iniciada	Passo que iniciou a etapa em questão do processo
Passo Final Etapa	String de caracteres	Valores categóricos em texto com o passo que finalizou aquela etapa do processo no sistema. Ex: 17 – Pendências Diversas regularizadas	Passo que encerrou a etapa em questão do processo

Tipologia_org	String de caracteres	Valores categóricos em texto com uma das 25 tipologias que classificam uma apólice de seguro garantia. Ex: Fiança Locatícia	Tipologia pertinente àquele processo
Conclusão	Data no formato dd/mm/aaaa	Valores de data entre 02/01/2015 e 07/06/2017.	Data de conclusão do processo.

Fonte: O autor (2017).

A partir desta lista de atributos, foram excluídos aqueles não pertinentes às análises da mineração, como por exemplo colunas existentes somente para cálculos na ferramenta Tableau ou colunas que apresentam sempre o mesmo valor (como a coluna “Status”, visto que todos os processos têm status “concluído” ou ainda colunas repetidas e de identificação interna da organização. Essa análise foi feita considerando apenas os aspectos gerais da base, mas não os aspectos técnicos das tarefas de mineração de dados e mineração de processos conforme detalhamento posterior.

Assim, foram retirados os atributos: filtro cíclico, status, data atualização, % atingimento, tempo líquido, segundos 18 horas, segundos 9,5 horas, segundos 8,5 horas, segundos inicial, segundos total, segundos final, proposta_doc e proposta. Com isso, restaram os atributos: código do CI, soma tempo líquido, tempo por etapa, etapa, área, emissor, tomador_doc, modalidade, inclusor, conclusor, data final, data inicial, tomador, passo encerramento, passo inicial etapa, passo final etapa, tipologia_org e conclusão.

3.3.2 Ferramentas

Neste trabalho, utilizou a ferramenta Tableau e a ferramenta Microsoft Excel para a análise estatística descritiva da base de dados trabalhada. O Tableau é um software comercial lançado em 2011 que tem versão paga e uma versão gratuita para estudantes. Permite análises de dados, geração de relatórios e painéis customizáveis de acordo com as bases de dados inseridas. A ferramenta trabalha com bases

estáticas e bases dinâmicas, atualizando de acordo com definições do usuário. Ela pode ser encontrada em https://www.tableau.com/pt-br através da versão paga ou as versões limitadas gratuitas.

Para a mineração de dados, utilizou-se a ferramenta Weka. O Weka é um software de mineração de dados *open source* desenvolvido em Java e distribuído gratuitamente. A ferramenta permite conduzir processos de mineração de dados de forma simples, com diversos algoritmos e variação de parâmetros. Além disso, a ferramenta oferece recursos para a execução de tarefas relacionadas ao pré-processamento de dados como, por exemplo, a seleção e a transformação de atributos. A ferramenta e mais informações podem ser encontradas em https://www.cs.waikato.ac.nz/ml/weka/.

Para a mineração de processos, utilizou-se a ferramenta Disco. O Disco é um software comercial lançado em 2012 pela Fluxicon e desenvolvido na linguagem Java. Seu diferencial é sua interface amigável e de fácil utilização. Uma característica relevante é que não há necessidade de decidir qual algoritmo de mineração deve ser usado, simplificando o uso e eliminando a diversidade de diferentes notações de modelos gerados pela mineração. Ele pode ser encontrado em https://fluxicon.com/disco/ e está disponível através de licenças pagas, licenças para estudantes e versões limitadas de avaliação.

3.4 PROCEDIMENTOS METODOLÓGICOS

Na Figura 9 são demonstrados os procedimentos metodológicos a serem realizados a fim de alcançar os objetivos específicos definidos nessa pesquisa, podendo assim, alcançar o objetivo geral.

FIGURA 9 - PROCEDIMENTOS METODOLÓGICOS

FONTE: O autor (2017)

Cada um dos procedimentos definidos será detalhado na sequência.

3.5 BASE DE DADOS

Como já mostrado na subseção 3.3.1, a base de dados utilizada traz dados de processos de subscrição recebidos e analisados pela organização. Contém mais de cento e cinquenta mil registros, incluindo aberturas de processos e suas diversas etapas. Em sua maioria, os atributos contidos na base trazem dados textuais, porém alguns dos principais estão em formato de data, de tempo ou até mesmo os dois.

A análise do formato dos dados é de suma importância para a escolha dos algoritmos utilizando na etapa de mineração de dados, visto que alguns algoritmos não funcionam bem com certos tipos de dados.

Para a mineração de processos, os atributos de tempo e data são os mais importantes pois indicam início, duração e fim das etapas dos processos.

3.6 ANÁLISE ESTATÍSTICA DESCRITIVA

A etapa de análise estatística descritiva é a etapa inicial da análise utilizada para descrever e resumir os dados. O objetivo desta análise é obter medidas estatísticas de distribuição dos valores existentes na base de dados para melhor entendimento dos atributos e identificação de aspectos que possam influenciar a aplicação dos métodos.

Algumas ferramentas do mercado possuem funcionalidades que permitem este tipo de análise. Entre as principais, destacam-se o SPSS e o Tableau. Justifica-se o destaque das duas pois o SPSS já foi utilizado em disciplina acadêmica e, portanto, já existe uma familiaridade com seu ambiente. O mesmo acontece com o Tableau, que já foi utilizado profissionalmente.

Abaixo está um quadro descritivo das duas ferramentas.

QUADRO 4 - FERRAMENTAS DE ANÁLISE ESTATÍSTICA

Software	**Descrição**	**Licença**
Tableau	O Tableau é um *software* comercial lançado em 2011 que tem versão paga e uma versão gratuita para estudantes. Permite análises de dados, geração de relatórios e painéis customizáveis de acordo com as bases de	Comercial/Licença para estudante

	dados inseridas. A ferramenta trabalha com bases estáticas e bases dinâmicas, atualizando de acordo com definições do usuário.	
SPSS	O Statistical Package for Social Science for Windows (SPSS) é um software para análise estatística de dados originalmente publicado em 1968. Permite inserção manual de dados e importação de bases em diversos formatos. O SPSS trabalha com duas perspectivas, a Data View (onde ocorre a entrada dos dados) e a perspectiva das variáveis, onde podemos selecionar nome, tipo, tamanho, rótulo, entre outros.	Comercial/Versão de Avaliação

FONTE: O AUTOR (2017)

Após as análises descritivas dos atributos, serão realizadas as etapas de mineração de dados e mineração de processos.

3.7 MINERAÇÃO DE DADOS

Na etapa de mineração de dados, serão escolhidas as tarefas de mineração e os algoritmos correspondentes a estas tarefas que melhor atendam aos requisitos da base de dados, de acordo com os atributos presentes na mesma. Os métodos e algoritmos definidos na revisão da literatura serão analisados e avaliados de modo a auxiliar na escolha do mais apropriado para extração de conhecimento da base de dados.

Essa etapa também envolve a escolha do *software* de mineração de dados a ser utilizado. Como critérios, primeiramente foram eliminados os *softwares* pagos. Por fim, optou-se pela utilização da ferramenta Weka.

O Weka é um *software* de mineração de dados *open source* desenvolvido em

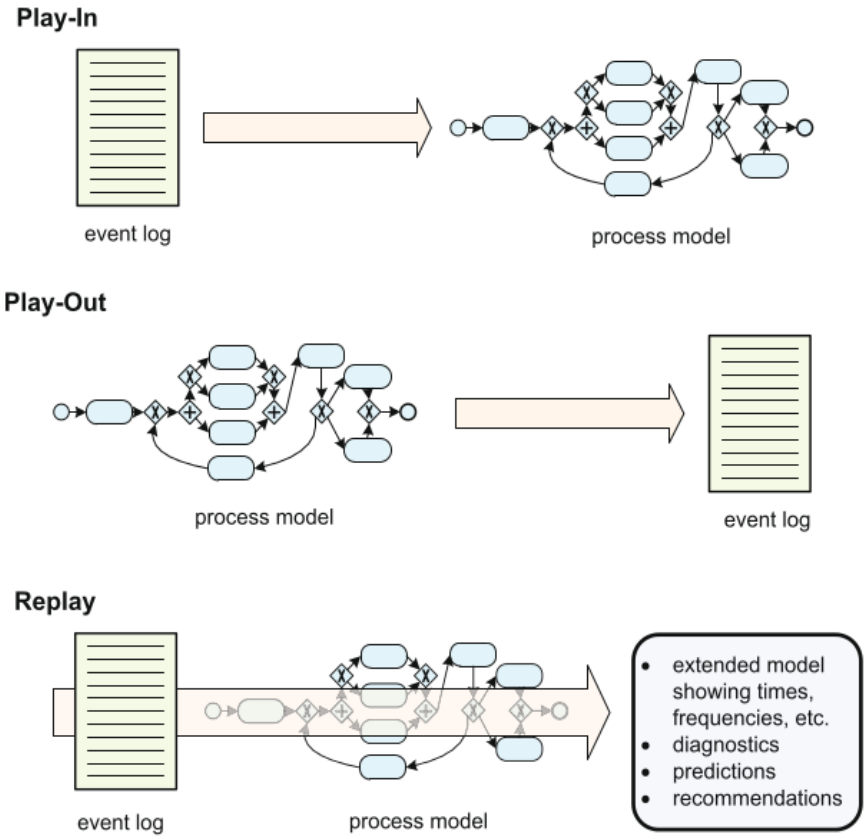
Java e distribuído gratuitamente. A ferramenta permite conduzir processos de mineração de dados de forma simples, com diversos algoritmos e variação de parâmetros. Além disso, a ferramenta oferece recursos para a execução de tarefas relacionadas ao pré-processamento de dados como, por exemplo, a seleção e a transformação de atributos.

A escolha dessa ferramenta se deu também pela familiaridade com o sistema, por já ter sido utilizado em disciplina acadêmica, apresentando resultados satisfatórios em termos de performance e abrangência de funções.

3.8 MINERAÇÃO DE PROCESSOS

Paralelamente a etapa de mineração de dados, ocorrerá a etapa de mineração de processos. Nesta, serão selecionadas as abordagens de mineração de processos mais adequadas à base de dados utilizada de forma a obter os melhores resultados. As abordagens definidas na revisão da literatura foram sintetizadas na Figura 10 para facilitar a visualização e análise e – enfim – auxiliar na escolha daquela mais apropriada.

FIGURA 10 - ABORDAGENS DE MINERAÇÃO DE PROCESSOS



FONTE: van der AALST (2017, p. 42).

Essa etapa também envolve a escolha do *software* de mineração de processos a ser utilizado. Dentre as ferramentas disponíveis no mercado, *ProM* e *Disco* são as mais conhecidas e possuem maior volume de documentação, facilitando seu uso. Sendo assim, se optará por uma das duas para a realização deste projeto. A descrição das duas ferramentas pode ser observada no Quadro 2.

QUADRO 5 - FERRAMENTAS PARA MINERAÇÃO DE PROCESSOS

Software	**Descrição**	**Licença**

ProM	Ferramenta gratuita que dispõe um ambiente integrado para mineração de processos a partir do logs de execução de processos. A ferramenta utiliza plug-ins que permitem flexibilidade durante o processo de mineração. É implementado na linguagem Java e possui código aberto, sendo possível os próprios usuários desenvolverem plug-ins para a ferramenta e também algoritmos de mineração de processos. A ferramenta ProM permite minerar diferentes perspectivas do log, como perspectivas de fluxo de dados, perspectiva organizacional, perspectiva de informação e perspectiva de aplicação.	Gratuita
Disco	Software comercial lançado em 2012 pela Fluxicon e desenvolvido na linguagem Java. Seu diferencial é sua interface amigável e de fácil utilização. Uma característica relevante é que não há necessidade de decidir qual algoritmo de mineração deve ser usado, simplificando o uso e eliminando a diversidade de diferentes notações de modelos gerados pela mineração.	Paga/Versão de Avaliação

FONTE: O autor (2017).

4 RESULTADOS E ANÁLISES

Nesta seção, são apresentadas a análise e descrição estatística da base de dados, a execução do algoritmo de mineração tradicional e mineração de processos e os respectivos resultados alcançados.

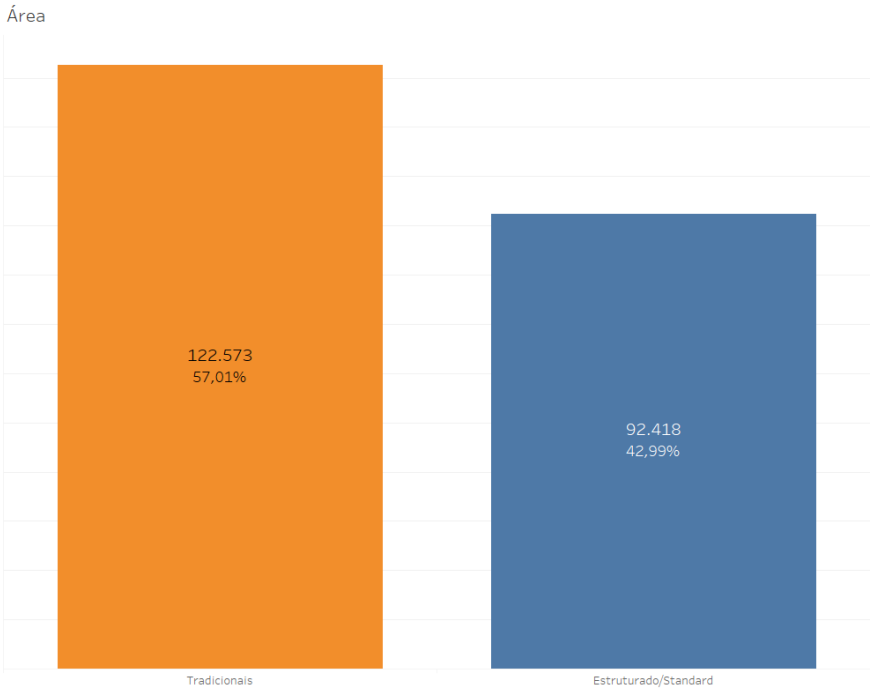
4.1 ESTATÍSTICA DESCRITIVA DA BASE

Inicialmente, realizou-se uma análise da base em termos da quantidade de

processo e etapas no total, assim como a variação do número de etapas no conjunto de processos trabalhados. 27.770 processos diferentes são encontrados na base, tendo 214.991 etapas no total. A média de etapas por processo é de 7,74 etapas. O maior número de etapas encontradas em um processo foi de 33 etapas, enquanto o menor foi de apenas uma etapa. 217 processos do conjunto total apresentam uma só etapa executada.

Para verificar a distribuição dos atributos categóricos na base de dados e identificar possíveis concentrações que pudessem influenciar no resultado final, foi realizada a contagem dos mesmos e posterior classificação em ordem decrescente através de gráfico de barras, tornando mais fácil a visualização daqueles que concentram a maior quantidade de registros da base de dados.

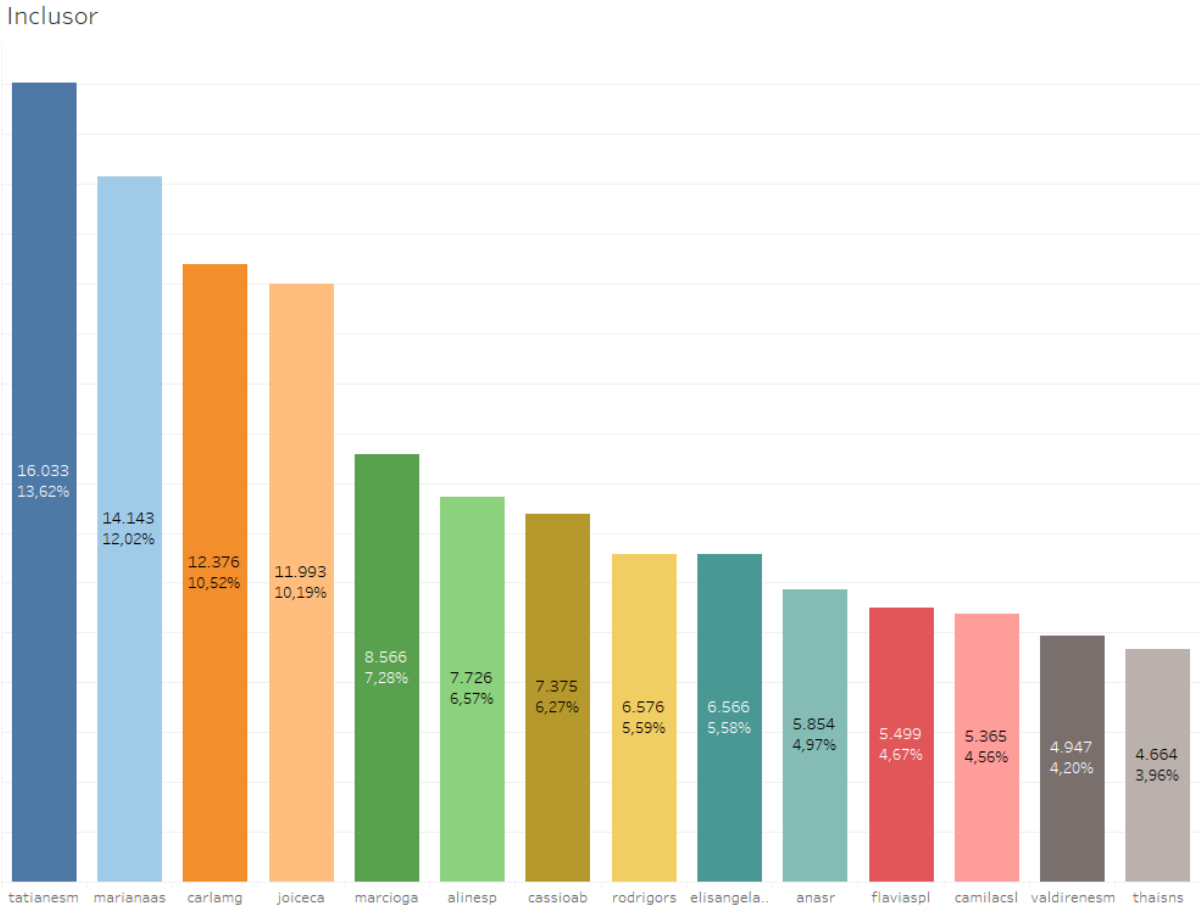
FIGURA 11 - DISTRIBUIÇÃO DOS REGISTROS POR “ÁREA”



FONTE: O AUTOR (2017).

A Figura 11 mostra que de todos os registros na base de dados, 122.573 (57,01%) foram da área “Tradicionais” e 92.418 (42,99%) foram da área “Estruturado/Standard”. Isso mostra que há uma predominância da área “Tradicionais”, porém sem que ela seja completamente dominante na base de dados.

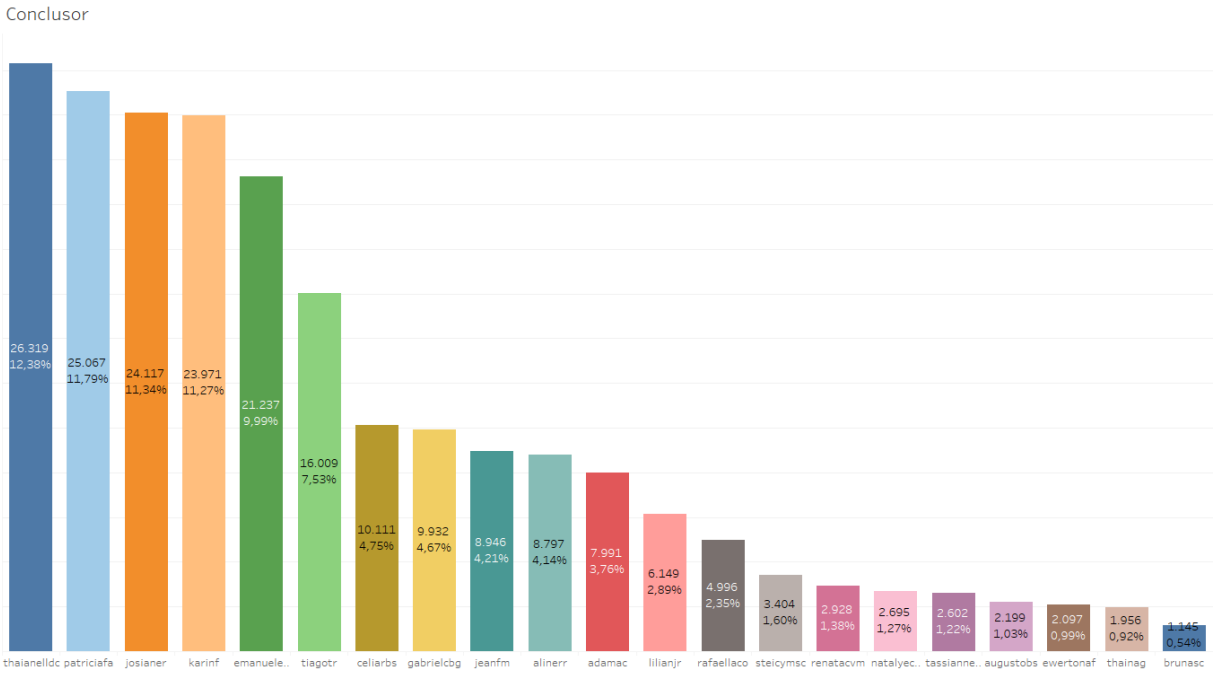
FIGURA 12 - DISTRIBUIÇÃO DOS REGISTROS POR “INCLUSOR”



FONTE: O AUTOR (2017).

O atributo “Inclusor” mostrado na Figura 12 indica o usuário que fez a inclusão da proposta de seguro no sistema. Na base trabalhada, ele apresenta 114 valores possíveis. Do total de registros, 53,63% estão concentrados em cinco inclusores (tatianesm, marianaas, carlamg, joiceca e marcioga) enquanto os outros 109 concentram apenas 46,37%.

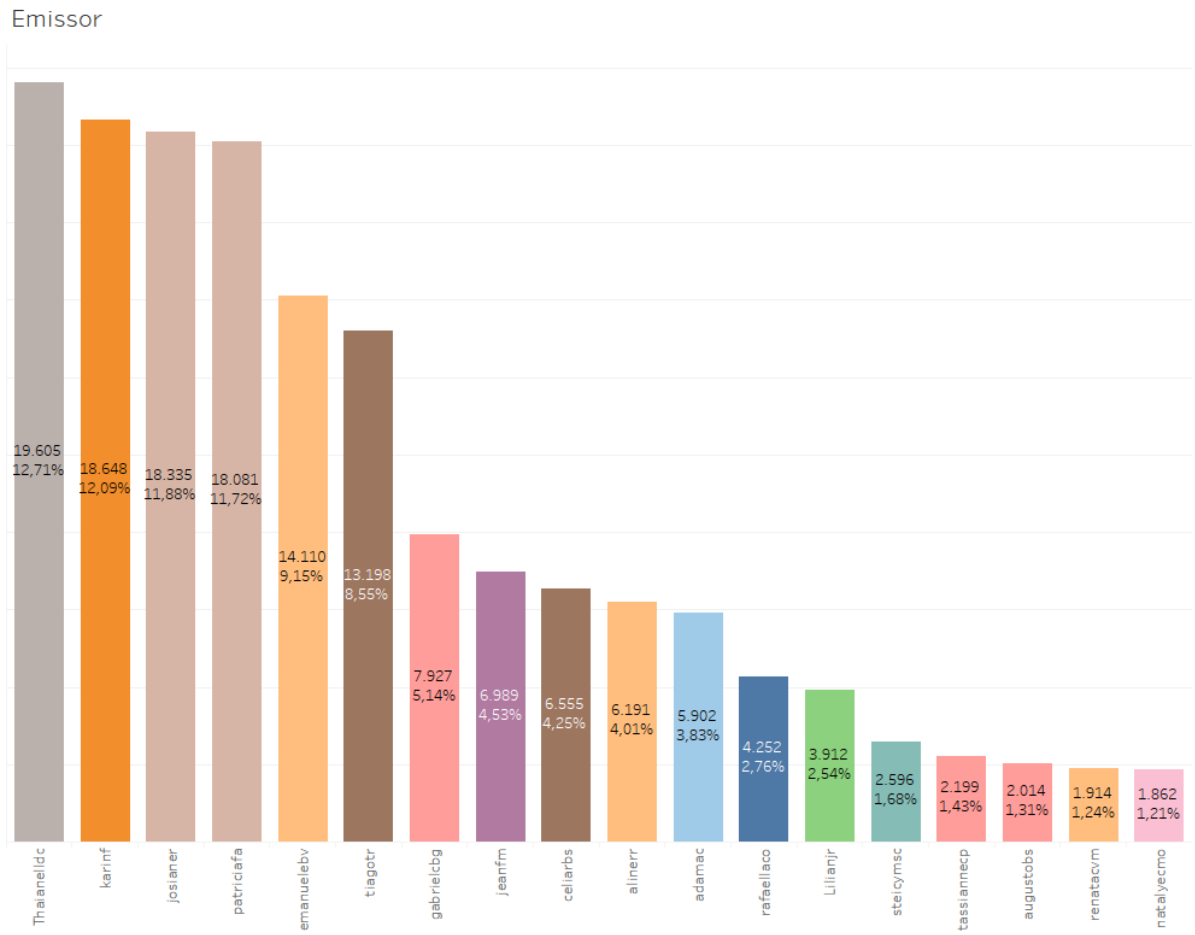
FIGURA 13 - DISTRIBUIÇÃO DOS REGISTROS POR “CONCLUSOR”.



FONTE: O AUTOR (2017).

O atributo “Conclusor” mostra quem foi o usuário que concluiu o processo no sistema. Ele apresenta 39 valores diferentes, porém a Figura 13 mostra que 64,3% de todos os registros estão concentrados nos usuários thaianelldc, patriciafa, josianer, karinf, emanuelebv e tiagotr. Todos os outros 36 somam apenas 35,7% dos processos.

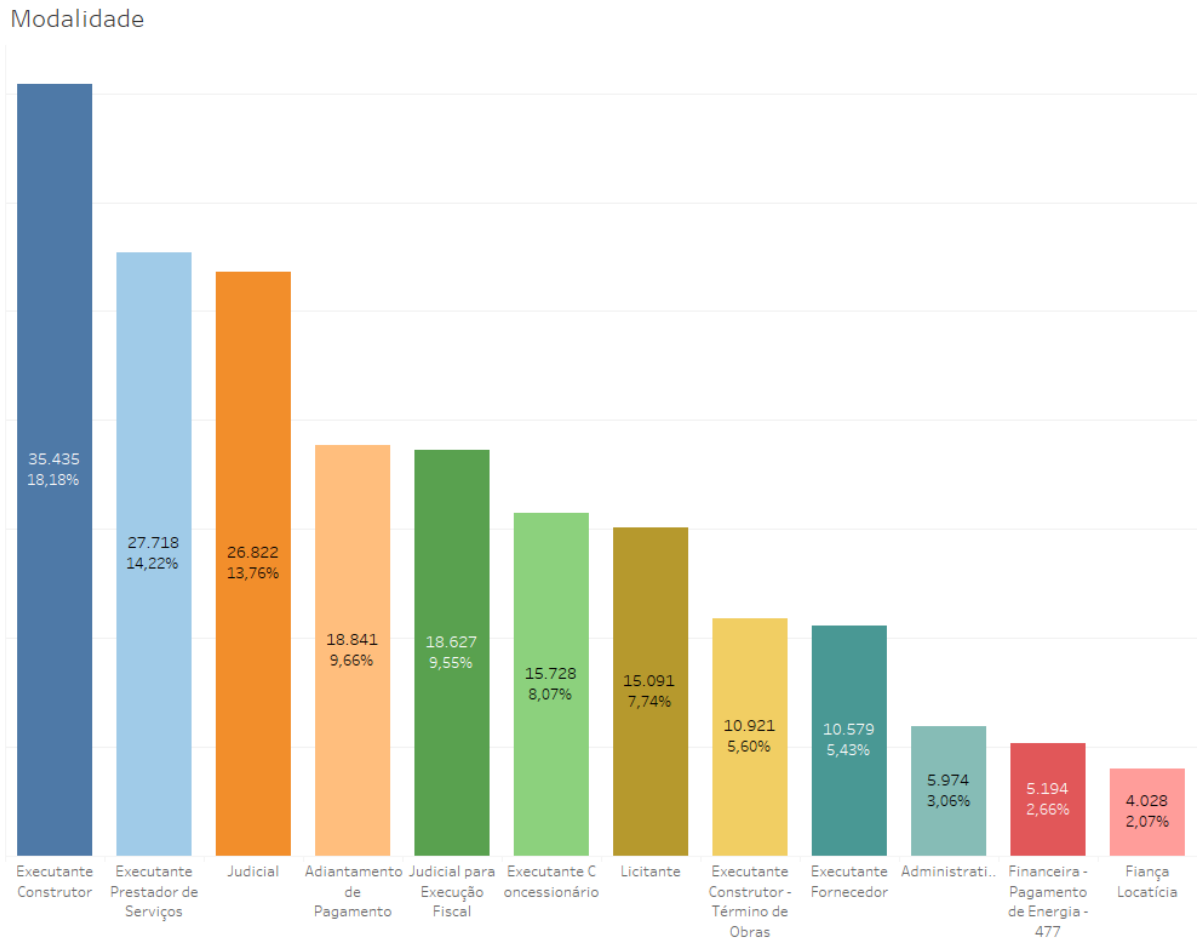
FIGURA 14 - DISTRIBUIÇÃO DOS REGISTROS POR “EMISSOR”



FONTE: O AUTOR (2017).

No caso do atributo “Emissor” (que mostra o usuário responsável pela emissão da apólice do seguro no sistema), também há uma concentração de 66,1% de todos os registros nos emissores thaianelldc, karinf, josianer, patriciaf, emanuelebv e tiagotr, como pode ser visto na Figura 14. Os outros 74 atributos somam apenas 33,9% dos registros da base.

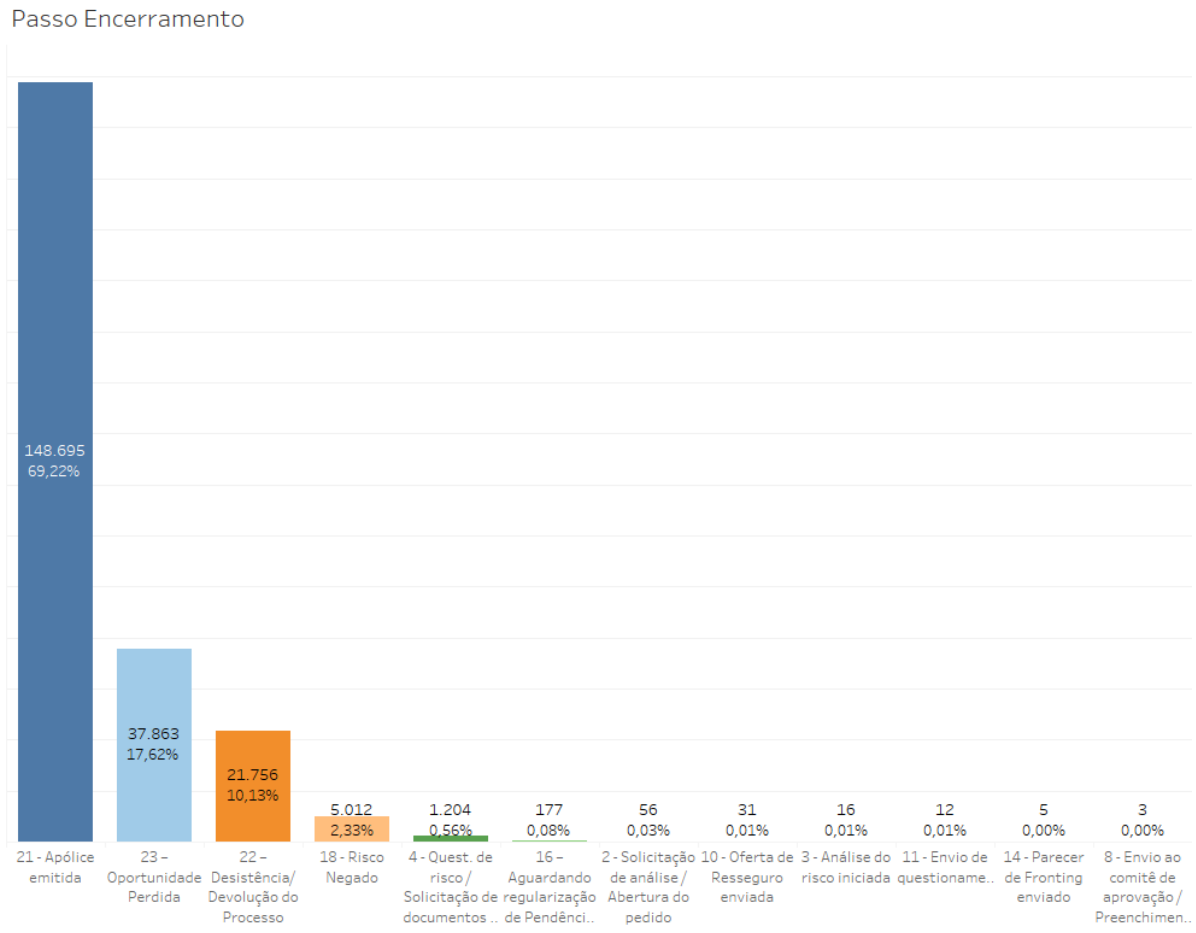
FIGURA 15 - DISTRIBUIÇÃO DOS REGISTROS POR “MODALIDADE”



FONTE: O AUTOR (2017).

O atributo “Modalidade” mostra a modalidade em que a apólice de seguro foi incluída no sistema. Ele apresenta 45 valores diferentes, porém a Figura 15 mostra que 65,37% de todos os registros estão concentrados em cinco modalidades: Executante Construtor, Executante Prestador de Serviços, Judicial, Adiantamento de Pagamento e Judicial para Execução Fiscal. Todos as outras 40 modalidades somam apenas 34,64% dos processos.

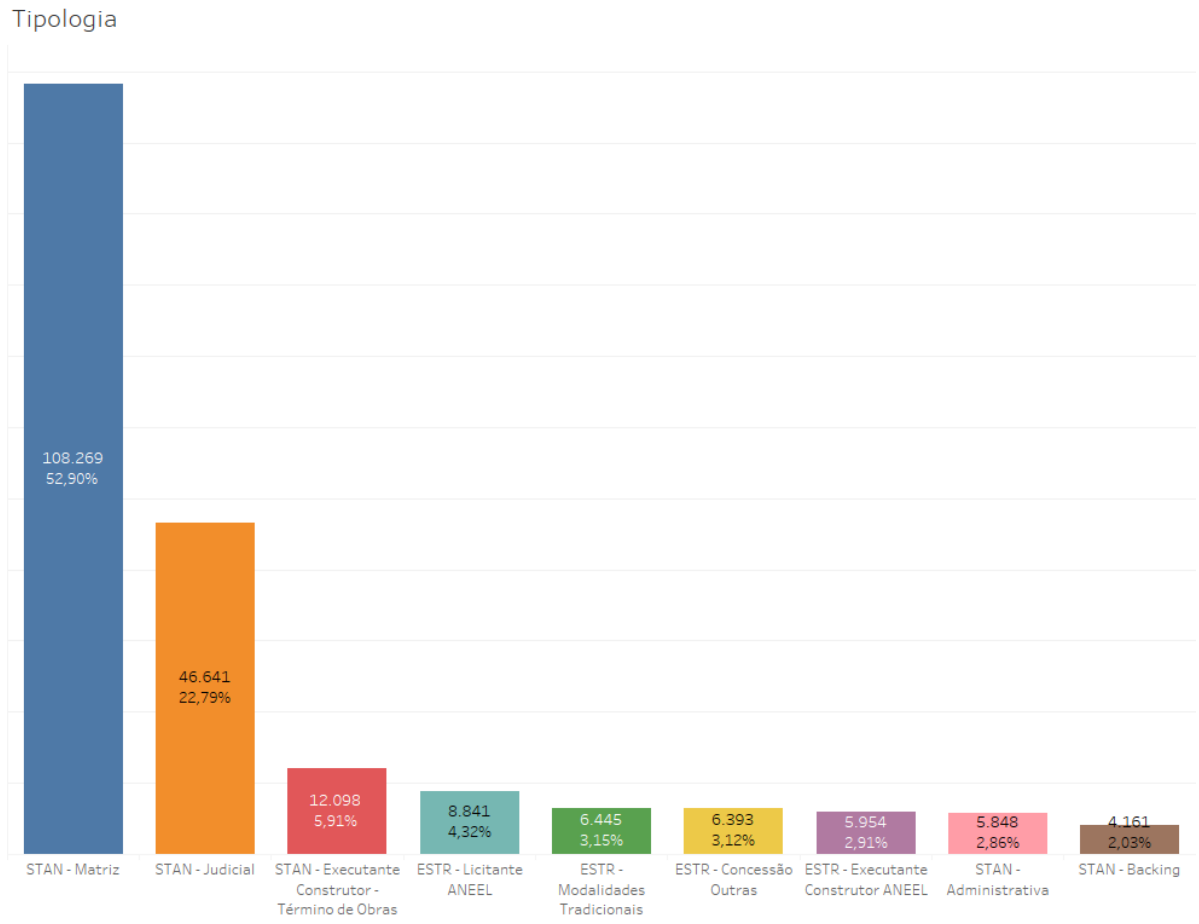
FIGURA 16 - DISTRIBUIÇÃO DOS REGISTROS POR “PASSO ENCERRAMENTO”



FONTE: O AUTOR (2017).

O atributo “Passo Encerramento” mostra qual o passo com que o processo foi finalizado. A análise da Figura 16 mostra que 86,84 de todos os registros da base contém apenas dois passos: Apólice emitida ou Oportunidade perdida. Apenas outros dois passos aparecem em mais de 1% dos casos, indicando que os outros oito passos existentes são pouquíssimos usados para encerrar os processos.

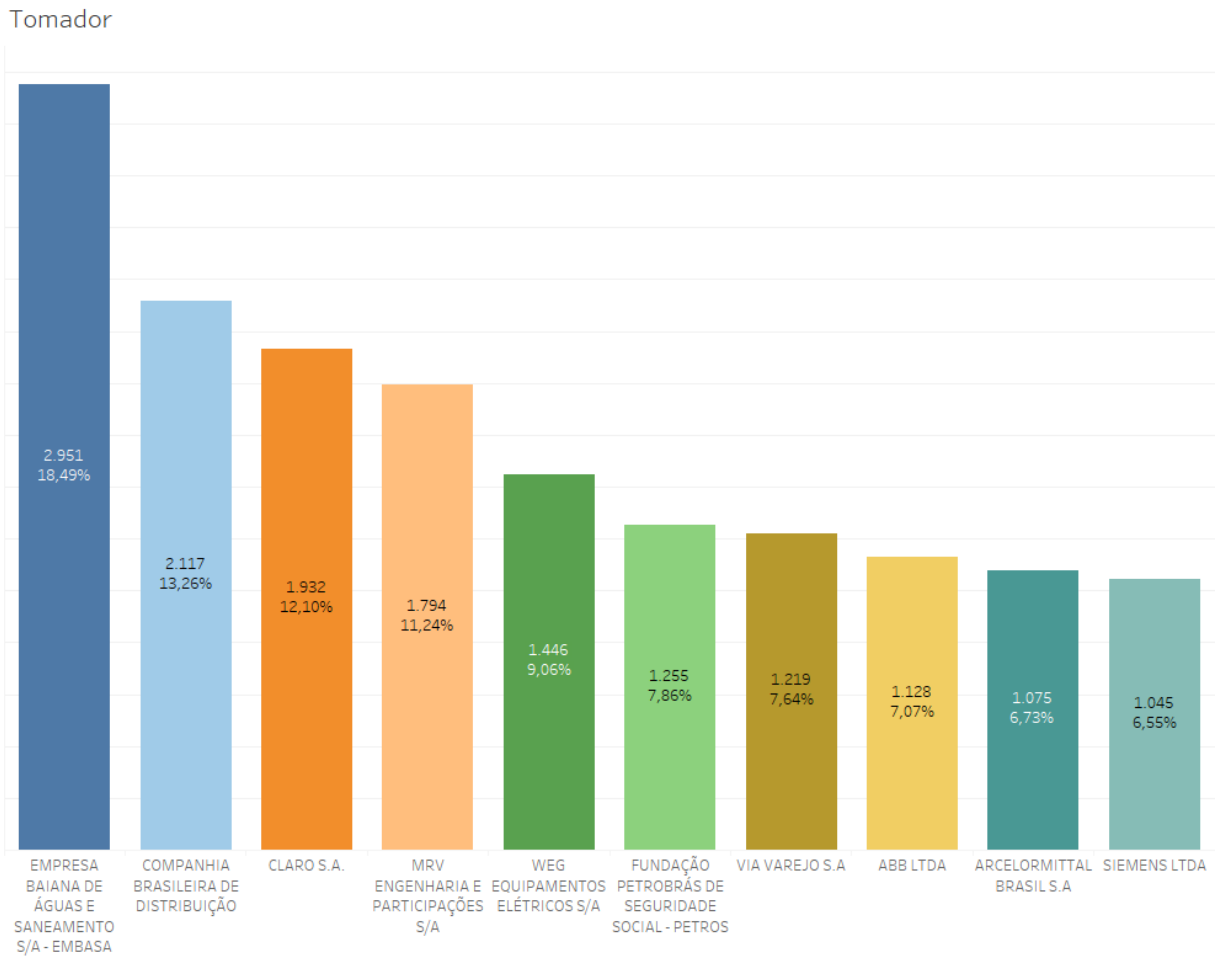
FIGURA 17 - DISTRIBUIÇÃO DOS REGISTROS POR “TIPOLOGIA”



FONTE: O AUTOR (2017).

O atributo “Tipologia” mostra com qual tipologia o processo foi classificado de acordo com as características do seguro. Na Figura 17, verifica-se que a tipologia STAN – Matriz concentra mais da metade dos registros (com 52,90%), enquanto a STAN – Judicial tem 22,79%, somando juntas 75,69%. As outras sete tipologias somam apenas 24,31%, menos de um quarto de todos os registros.

FIGURA 18 - DISTRIBUIÇÃO DOS REGISTROS POR “TOMADOR”



FONTE: O AUTOR (2017).

Por fim, o atributo “Tomador” traz a razão social da empresa responsável pelas obrigações da apólice do seguro. Este atributo possui mais de sete mil valores diferentes, porém 99,99% dos casos estão concentrados em apenas 10 Tomadores.

Ao concluir a análise e descrição estatística da base de dados, a subseção seguinte aborda a mineração de dados e os ajustes realizados nesta etapa do trabalho.

4.2 MINERAÇÃO DE DADOS

Para a utilização da base na mineração de dados, a coluna Código do CI foi retirada, visto que o ID do processo não traz qualquer ganho de informação à esta modalidade de mineração. Esse é o primeiro ponto onde a mineração de dados tradicional se mostra pouco hábil para lidar com processos, visto que sem a identificação do código a que cada etapa pertence, é impossível caracterizar os processos e considerar suas diferentes etapas como um conjunto, ou seja, o fluxo do processo é ignorado na mineração de dados tradicional.

Em seguida, considerando o fato de os principais algoritmos de mineração não trabalharem bem com atributos no formato de tempo ou data, os atributos que se encontravam neste formato foram discretizados conforme descrito abaixo.

O único atributo presente base no formato de tempo foi o “soma tempo líquido”, e para ele foi utilizada a discretização em faixas de tempo. Levando em conta as análises de tempo realizadas na organização e para um maior nível de detalhe e consequentemente maior nível de informação, foram utilizadas cinco faixas de tempo: até 10 minutos; de 10 a 20 minutos; de. 20 a 30 minutos; de 30 minutos a 1 hora; de 1 a 2 horas; mais de 2 horas. Os limites de cada faixa podem ser vistos no Quadro 6. Para “soma tempo líquido”, foi obtido o valor máximo de 136:33:11 e o valor mínimo de 00:00:00.

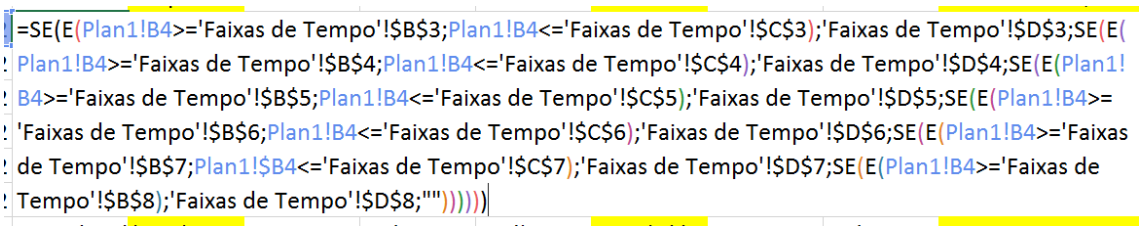
QUADRO 6 - FAIXAS DE TEMPO PARA "SOMA TEMPO LÍQUIDO"

Faixas de Tempo		
0:00:00	0:09:59	Até 10 minutos
0:10:00	0:19:59	De 10 a 20 minutos
0:20:00	0:29:59	De 20 a 30 minutos
0:30:00	0:59:59	De 30 minutos a 1 hora
1:00:00	1:59:59	De 1 a 2 horas
2:00:00	-	Mais de 2 horas

FONTE: O AUTOR (2017).

A discretização foi realizada na ferramenta Microsoft Excel pela utilização de uma fórmula que comparou os tempos do atributo “soma tempo líquido” com os limites inferiores e superiores de cada faixa de tempo. A fórmula pode ser observada na Figura 19.

FIGURA 19 - DISCRETIZAÇÃO DO ATRIBUTO "SOMA TEMPO LÍQUIDO"



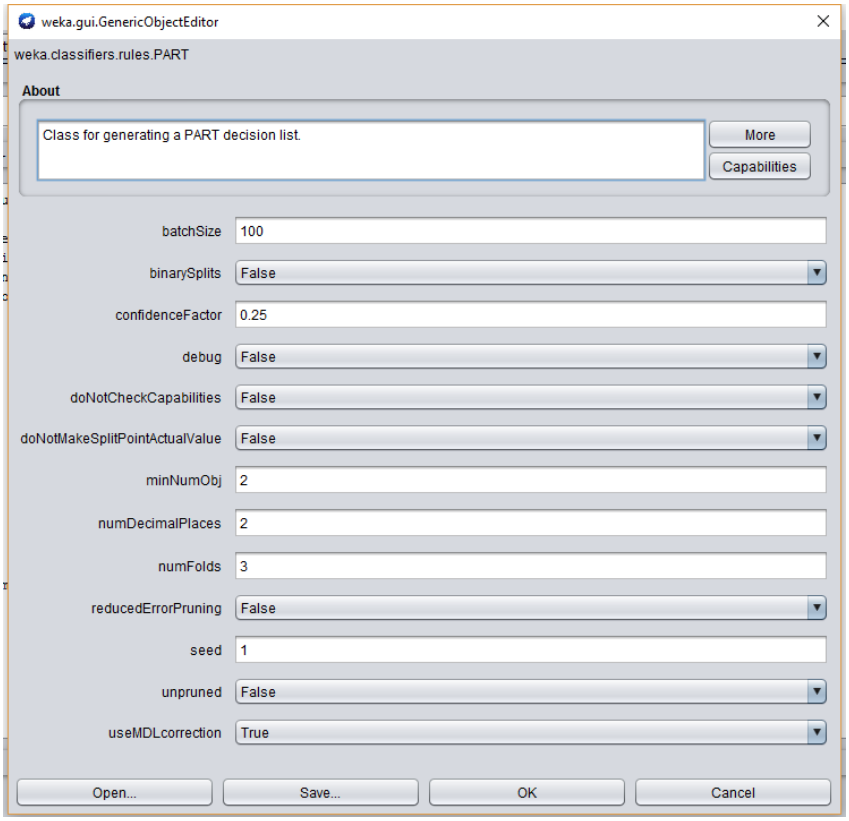
FONTE: O AUTOR (2017).

Após a discretização, a base foi salva no formato *csv* e posteriormente convertida para o formato *arff* (formato de entrada da ferramenta Weka). A conversão de *csv* para *arff* foi realizada em conversor disponível na própria ferramenta.

Na primeira tentativa de realização da mineração, a ferramenta não conseguiu carregar a base utilizada, informando que havia um erro no código *arff*. Após analisar a base, verificou-se que o problema estava na coluna “tomador”, pois a mesma possuía um número muito grande de caracteres e também caracteres especiais, impossibilitando a ferramenta de trabalhar com este atributo. Assim, ele foi retirado da base e o processão de conversão de *csv* para *arff* teve que ser refeito.

Após essa correção, a base foi carrega no *Weka* e optou-se por utilizar o algoritmo de classificação PART, que é uma variação do algoritmo J48 descrito anteriormente na subseção 2.2.2.1. A execução foi feita com os parâmetros padrões do Weka para o PART (mostrado na Figura 20) e levou 668,5 segundos para ser executado.

FIGURA 20 - PARÂMETROS DE EXECUÇÃO DO ALGORITMO PART



FONTE: O AUTOR (2017).

O atributo meta definido foi o passo de conclusão de processo, ou seja, o algoritmo tenta inferir qual será a conclusão do processo através da análise dos atributos. Os resultados obtidos podem ser vistos na Figura 21, contendo as quatro regras com maior ganho de informação geradas pelo algoritmo (as demais regras apresentaram ganho de informação muito abaixo destas quatro).

FIGURA 21 - RESULTADOS PARA O ALGORITMO PART

```
=== Run information ===

Scheme:       weka.classifiers.rules.PART -M 2 -C 0.25 -Q 1
Relation:     Base DM limpa-weka.filters.unsupervised.attribute.Remove-R2
Instances:    214989
Attributes:   13
              Faixa de Tempo Soma Tempo Líquido
              Etapa
              Área
              Emissor
              Modalidade
              Inclusor
              Conclusor
              Data Inicial
              Data Final
              Passo Encerramento
              Passo Inicial Etapa
              Passo Final Etapa
              Tipologia_org
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

PART decision list
------------------

Passo Inicial Etapa = 21 - Apólice emitida: 21 - Apólice emitida (12193.0/1.0)

Passo Inicial Etapa = 23 – Oportunidade Perdida: 23 – Oportunidade Perdida (3163.0)

Passo Inicial Etapa = 22 – Desistência/Devolução do Processo: 22 – Desistência/Devolução do Processo (2402.0)

Passo Final Etapa = 21 - Apólice emitida: 21 - Apólice emitida (6178.0/1.0)

Conclusor = brunasc AND
Inclusor = brunopds: 21 - Apólice emitida (176.0)
```

FONTE: O AUTOR (2017).

Ao analisar as quatro melhores regras obtidas pelo PART, concluiu-se que as mesmas não trazem conhecimento relevante para o entendimento dos processos analisados e suas etapas. Como já dito anteriormente, isso se dá pelo fato de os métodos mineração tradicionais não conseguirem trabalhar processos considerando seu fluxo como um todo, apenas etapas separadas uma por uma.

Assim, considerando-se que os resultados da mineração de dados não se mostraram efetivos, optou-se pela utilização de ferramenta adequada para a realização da mineração de processos, conforme apresentado na seção 4.3.

4.3 MINERAÇÃO DE PROCESSOS

Visando melhores resultados que a mineração de dados tradicional na extração de conhecimento da base trabalhada, utilizou-se a mineração de processos. Ela foi realizada utilizando a ferramenta Disco.

Para esta modalidade, não foi necessária a exclusão manual de colunas não utilizadas pois a ferramenta possui uma função de classificação dos atributos onde é possível defini-los como o código identificador de um processo, uma atividade (ou etapa) de processo, um indicador de data e hora, um recurso utilizado no processo, outros tipos de atributos ou ainda inutilizá-lo na execução do algoritmo de mineração. Essas classificações podem ser vistas na Figura 22.

FIGURA 22 - CLASSIFICAÇÕES POSSÍVEIS PARA UM ATRIBUTO NO DISCO

FONTE: Fluxicon Disco (2017).

A coluna “Código do CI” foi utilizada para identificar os diferentes processos e suas respectivas atividades; as colunas “Inclusor”, “Emissor” e “Conclusor” foram utilizadas como recursos dos processos; as colunas “Data Inicial” e “Data Final” foram utilizadas como *timestamps* indicando o início e o fim das etapas e do processo; numa primeira tentativa, as colunas “Passo Inicial Etapa” e “Passo Final Etapa” foram utilizadas como atividades do processos (para indicar em que passo cada etapa começou e terminou”, porém isso levou a um volume muito de grande de texto que dificultou a leitura do modelo gerado e então a coluna “Passo Final Etapa” foi retirada, deixando apenas a coluna “Passo Inicial Etapa” como identificador das etapas; as colunas “Faixa de Tempo Soma Tempo Líquido”, “Faixa de Tempo por Etapa”, “Área”, “Modalidade” e “Tipologia” foram utilizadas como informações adicionais ao funcionamento dos processos (colunas classificadas como “outros” na ferramenta).

O algoritmo utilizado nesta modalidade de mineração é um algoritmo desenvolvido por Christian W. Gunther base na lógica fuzzy e denominado *Fuzzy Miner*. Ele funciona gerando um modelo de processo através de um *log* inserido na ferramenta, neste caso, a base de processos extraída na organização. Este modelo é simplificado usando agregações de atividades/etapas de comportamento similares dentro do processo, abstração de atividades com pouca relevância e ênfase naquelas de maior relevância. Essa relevância é determinada através de fatores como

frequência com que as atividades ocorrem, correlação entre atividades e a ordem em que elas acontecem.

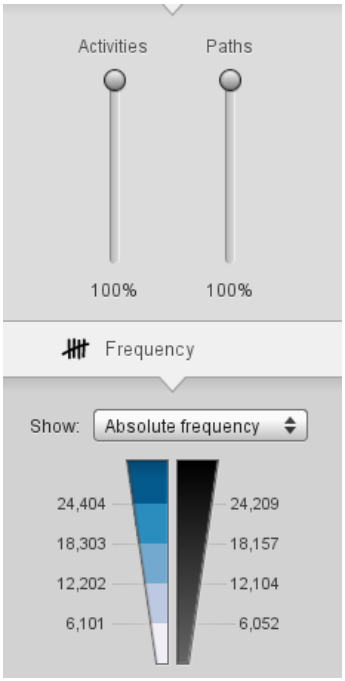
O modelo gerado com a execução do algoritmo depende do ajuste dos parâmetros existentes na ferramenta. Estes parâmetros e seus possíveis valores estão listados abaixo:

- *Activities*: de 0% a 100%. Indica quais atividades devem ser mostradas no modelo de acordo com a sua frequência de ocorrência. Exempo: Se o valor escolhido para este parâmetro for 100%, o modelo mostra todas as atividades executadas em todos os casos. Em 0%, mostra apenas aquelas mais frequentes;
- *Paths*: de 0% a 100%. Indica quais fluxos (ou caminhos) devem ser mostradas no modelo de acordo com a sua frequência de ocorrência. Exemplo: Se o valor escolhido para este parâmetro for 100%, o modelo mostra todas os fluxos executadas em todos os casos. Em 0%, mostra apenas os fluxos mais frequentes, ou seja, o fluxo mais executado para aquele processo.

O algoritmo apresenta duas possibilidades de seleção de “métrica meta” para usar como base na geração do modelo: frequência de ocorrência ou duração. Quando utilizada a frequência, pode-se optar pela frequência absoluta de uma atividade no universo de processos ocorridos ou o número máximo de repetições que esta atividade teve nos processos executados. Quando utilizada a duração da etapa, pode-se optar pela duração máxima das atividades, a duração média, a duração máxima ou a duração mínima. Para este trabalho, optou-se por trabalhar apenas considerando a frequência como métrica, visto que o objetivo está em analisar os diferentes fluxos seguidos pelo processo e não sua duração.

Primeiramente, foi realizada a execução do algoritmo com os parâmetros definidos de modo a mostrar todas as atividades realizadas e todos os fluxos seguidos pelo processo, tendo como métrica a frequência absoluta das atividades. Estes parâmetros podem ser observados na Figura 23.

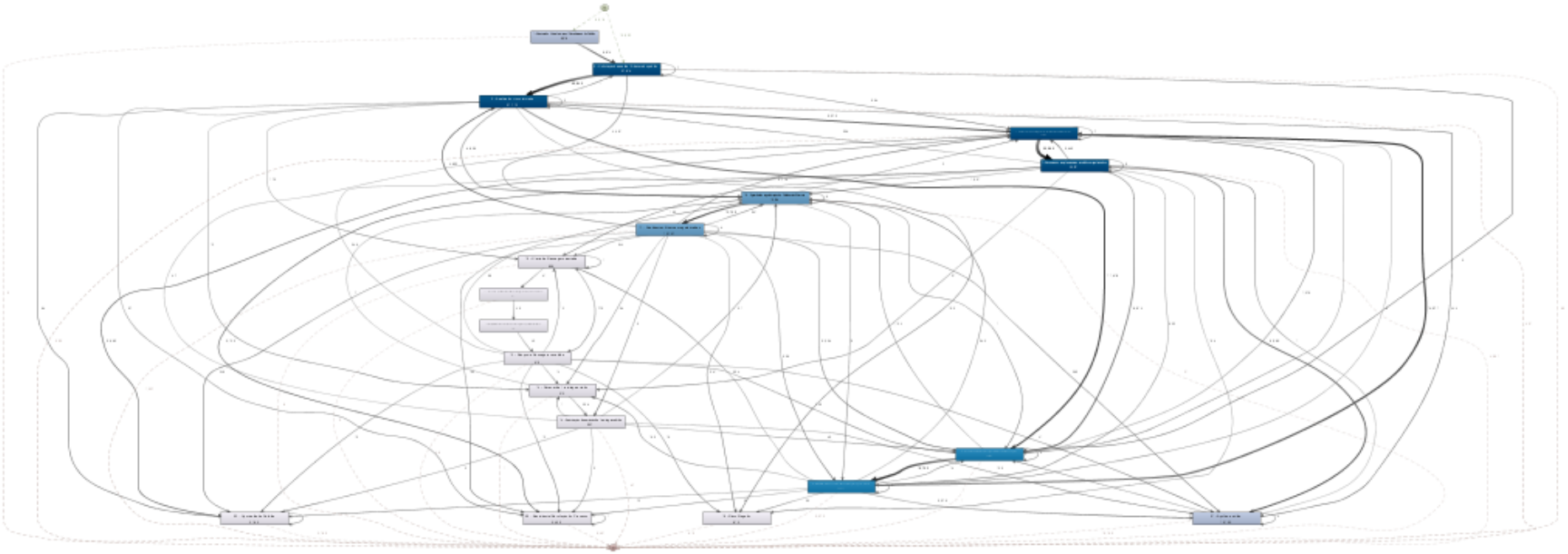
FIGURA 23 - PARÂMETROS DA PRIMEIRA EXECUÇÃO NO DISCO



FONTE: O AUTOR (2017).

O modelo obtido com estes parâmetros é apresentado na Figura 24 e é um tipo de modelo chamado de “spaghetti” por apresentar um número muito grande de fluxos e lembrar um emaranhado de macarrões, o que dificulta a análise pela dificuldade de visualização. Este tipo de modelo aparenta ser o mais indicado para uma análise minuciosa do comportamento do processo e consequentemente uma possibilidade maior de descoberta de novos conhecimentos acerca do mesmo. No entanto, para isso acontecer, é necessário filtrar fluxos específicos do processo para melhorar a visualização dos mesmos, o que exige maior tempo e atenção para analisar o grande volume de fluxos a se analisar.

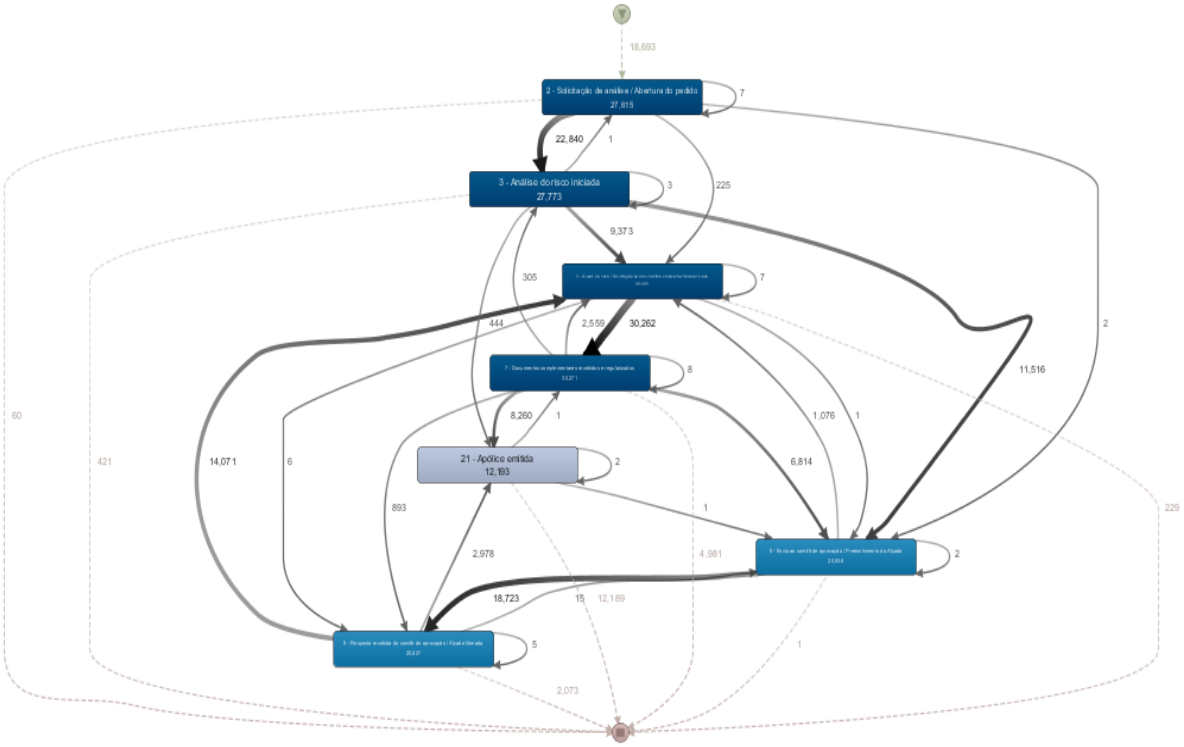
FIGURA 24 - MODELO GERADO NA PRIMEIRA EXECUÇÃO



FONTE: O AUTOR (2017).

Em seguida, o algoritmo foi executado uma segunda vez com os parâmetros definidos de modo a mostrar apenas as atividades mais frequentes (*Activities* = 0%) mantendo todos os fluxos realizados nos processos (*Paths* = 100%). O modelo gerado com estes parâmetros é mostrado na Figura 25.

FIGURA 25 - MODELO GERADO NA SEGUNDA EXECUÇÃO

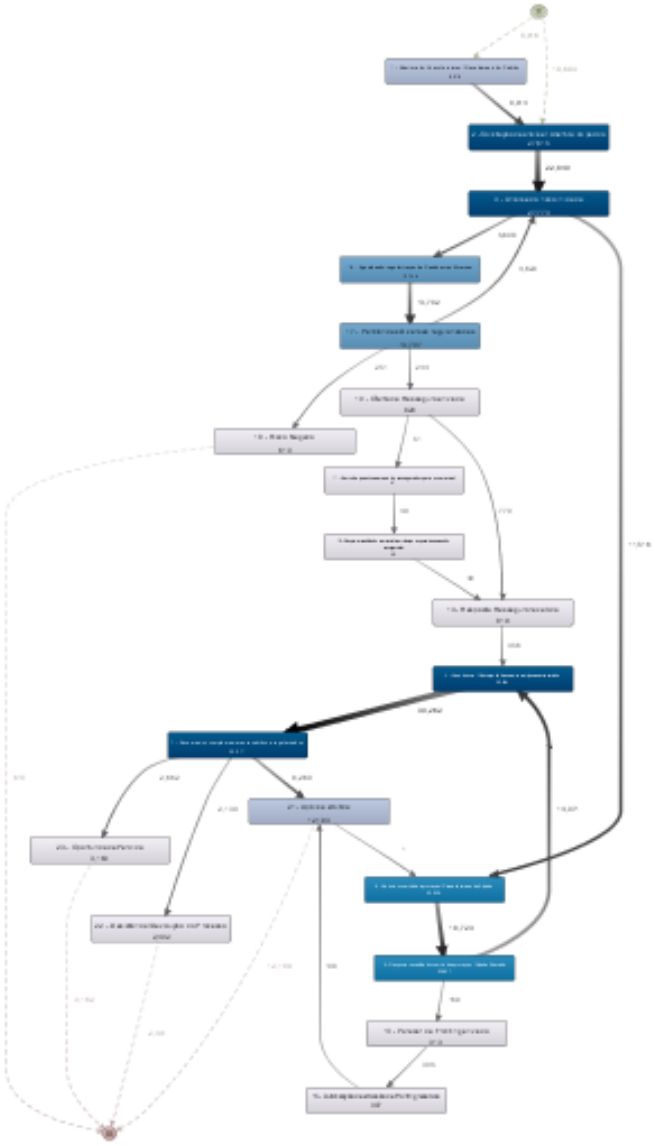


FONTE: O AUTOR (2017).

Como este modelo foca principalmente nos fluxos seguidos e deixa de lado atividades pouco frequentes, se mostra o mais indicado para a análise deste trabalho, onde a busca é por descoberta de conhecimento relacionado ao comportamento do processo e seus diversos caminhos.

Uma terceira execução foi realizada mostrando todas as atividades realizadas nos processos (*Activities* = 100%) e apenas os fluxos mais frequentes (*Paths* = 0%), conforme demonstra a Figura 26.

FIGURA 26 - MODELO GERADO NA TERCEIRA EXECUÇÃO

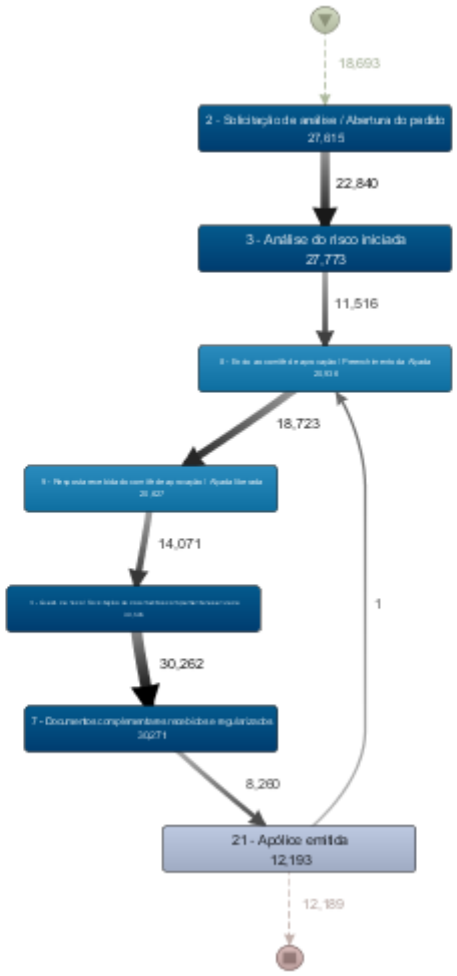


FONTE: O AUTOR (2017).

Este modelo mostra todas as atividades executadas no universo do processo analisado, mesmo aquelas menos frequentes (até mesmo se for executada uma única vez).

A quarta execução mostra apenas as atividades mais frequentes (*Activities* = 0%), assim como os fluxos mais frequentes (*Paths* = 0%) e pode ser vista na Figura 27.

FIGURA 27 - MODELO GERADO NA QUARTA EXECUÇÃO



Fonte: O autor (2017).

Este modelo é o mais simplificado possível e mostra somente os aspectos mais frequentes do processo executado. Pode se dizer que este é o modelo ideal do processo, onde não há nenhum desvio do fluxo principal e as etapas são seguidas sempre na mesma ordem.

Os quatro modelos gerados tiveram como valores para seus parâmetros apenas 0% ou 100% para a simplificação da análise, porém é possível variar os mesmos no caso de uma análise que varia seu foco entre as diferentes atividades ou os diferentes fluxos existentes.

Como já dito, o modelo gerado na segunda execução (mostrado na Figura X) se mostrou o mais pertinente para a análise neste trabalho por focar nos diferentes fluxos seguidos pelo processo. A partir deste modelo, é possível verificar que diversos processos entram em um *loop*, repetindo a mesma etapa diversas vezes, algo que não deveria acontecer levando em consideração o contexto da organização e seu

sistema. Também é possível ver uma variedade grande no fluxo de finalização do processo, onde há sete passos diferentes atuando como finalizadores (de um total de 23 etapas possíveis) e o passo 21 – Apólice emitida” é o mais frequente, com 12.189 processos.

Analisando os fluxos presentes, verifica-se que o passo 2 é o mais utilizado após o passo 1 (18.693 casos, ou 67% do total) e o passo 3 logo após esses dois (22.840 dos casos, ou 82% do total).

Este modelo também demonstra que os fluxos que passam pelo passo “7 – Documentos complementares recebidos e regularizados” possuem chances de resultar no encerramento do processo, visto que de um total de 30.271 ocorrências dele, 4.891 casos se encerram (mesmo podendo ainda passar por outro passo antes de isso acontecer de fato). Outro passo onde ocorre algo semelhante é o passo “9 – Resposta recebida do comitê de aprovação/Alçada liberada”, com 2.073 casos resultando em encerramento (o terceiro mais relevante no modelo gerado).

4.4 ANÁLISE DOS RESULTADOS

A partir da análise estatística da base de dados, foi possível verificar que o atributo "Passo Encerramento" apresentou a maior concentração entre os atributos nominais, com 69,22% de ocorrência de apenas um valor entre os vinte e três valores possíveis. O atributo "Tipologia" também apresentou alta concentração, com 52,90% de concentração em "STAN - Matriz" e 22,79% em "STAN - Judicial". Os demais atributos apresentaram, no geral, altas concentrações em cinco valores do universo possível daqueles atributos. Esta análise possibilitou levar isso em conta ao avaliar os resultados da mineração de dados e mineração de processos.

Na mineração de dados, os resultados obtidos com a execução do algoritmo PART na ferramenta Weka se mostraram pouco satisfatórios, pois o algoritmo não conseguiu trabalhar com a ideia de diversas etapas dos processos interligadas apenas pelo ID do processo. Apesar de o algoritmo ter sido executado sem maiores problemas após o tratamento da base de dados, as regras geradas (vistas na Figura 21) apresentaram pouca importância para o contexto da pesquisa.

Na mineração de processos, a ferramenta Disco e o algoritmo utilizado se mostraram extremamente eficientes para trabalhar com bases de dados de processos, possibilitando uma visão abstraída dos fluxos existentes, eliminando detalhes

irrelevantes e destacando aquilo de maior importância considerando a base de dados como um todo.

Uma grande variação de modelos gerados foi possível através do ajuste dos parâmetros do algoritmo utilizado. Quatro modelos diferentes foram gerados com enfoques específicos utilizando valores binários para os parâmetros (0 ou 100), sendo o segundo modelo o mais relevante para o contexto deste trabalho, pois focou na análise dos diferentes fluxos que acontecem no universo de processos trabalhados. Porém, outros tipos de modelos podem ser gerados variando os parâmetros de outras maneiras e para serem aplicados em contextos diferentes, dependendo da análise que se quer realizar.

Deste modo, fica óbvio que a mineração de processos é a modalidade mais indicada para trabalhar com dados de processo operacionais, visto que a mineração de dados tradicional não conseguiu trabalhar de maneira satisfatória com este modelo de base de dados.

5 CONSIDERAÇÕES FINAIS

O interesse na ciência de dados vem crescendo rapidamente. Isto acontece devido ao volume cada vez maior de dados gerados devido a evolução da tecnologia e do uso de sistemas de informação cada vez mais robustos que permite a coleta de dados de maneira mais completa e eficiente. Dentro do ambiente organizacional, este interesse também se deve a busca de diferentes formas de extrair informação e conhecimento, pois as bases de dados cresceram tanto que dificultaram a análise manual, sendo necessária a aplicação de diferentes técnicas e ferramentas capazes de analisar estas bases de forma eficiente para alcançar resultados interessantes. Isto inclui a análise de processos operacionais nas organizações.

Neste contexto, a mineração de dados e a mineração de processos entram com abordagens diferentes auxiliar o trabalho de gestão da informação nas organizações. Diante das diferenças nas duas abordagens, são necessárias escolhas que se encaixem nos modos de execução esperados, como escolha das tarefas, dos algoritmos, abordagens e ferramentas.

Também é importante a análise dos atributos presentes nas bases de dados e seus formatos. Assim, para ambas minerações de dados e de processos, são importantes as etapas de preparação dos dados antes de sua aplicação - realizando a limpeza dos dados e a remoção de ruídos para garantir a qualidade e a eficiência dos algoritmos aplicados. Faz parte desta preparação também a análise estatística descritiva da base e seus atributos para melhor entendimento e identificação de aspectos que possam influenciar a aplicação dos métodos.

5.1 ALCANCE DOS OBJETIVOS

O objetivo geral deste trabalho foi aplicar técnicas de mineração de dados e mineração de processos em uma base de dados de seguro garantia, de modo a identificar padrões em processos de subscrição desta modalidade de seguro. Para isto, buscou-se alcançar quatro objetivos específicos.

O primeiro objetivo específico foi definido como estudar e definir, dentre os métodos mais citados na literatura científica, quais os mais adequados para a mineração de dados da base em questão. Este objetivo foi alcançado através do levantamento bibliográfico realizado de modo a identificar os principais métodos

utilizados na literatura e que apresentassem maior capacidade de lidar com o modelo da base de dados trabalhada e que apresentasse resultados de fácil entendimento e visualização para o contexto em questão. Seguinte ao levantamento teórico, optou-se pela tarefa de classificação para realização da mineração de dados tradicional, considerando que a mesma é a mais indicada para alcançar o resultado esperado de classificar os resultados dos processos de acordo com o comportamento em suas etapas. Em seguida, foi selecionado o algoritmo a ser utilizado na mineração dentre as opções habilitadas na ferramenta Weka levando em conta as características da base e seus atributos.

O algoritmo selecionado para a classificação foi o PART, visto que o mesmo trabalha com resultados em formatos de regras, estabelecendo uma ligação entre os atributos presentes e o resultado final para o atributo meta definido. Entretanto, os resultados se mostraram pouco satisfatórios, pois o algoritmo não conseguiu trabalhar bem com as diversas etapas dos processos divididas uma em cada registro da base. Assim, as regras geradas apresentaram pouca importância para o contexto da pesquisa.

O segundo objetivo específico (definir uma ou mais ferramentas de mineração de processos a ser(em) utilizadas neste projeto) foi atingido após avaliar quais das ferramentas disponíveis melhor atendiam as necessidades para este projeto. Duas ferramentas foram avaliadas – ProM e Disco – e a última se mostrou ser a melhor escolha devido à sua maior facilidade de uso e interface amigável, assim como a qualidade superior do modelo gerado. A ferramenta Disco possui a capacidade customização dos resultados obtidos através da variação dos parâmetros do algoritmo de mineração de processos e isso tem um grande impacto nas possíveis análises a serem feitas utilizando-a.

O terceiro objetivo específico (preparar a base para descoberta de padrões em conformidade com os métodos e ferramentas escolhidos) foi alcançado de forma diferente para a mineração de dados e para a mineração de processos. Na mineração de dados, foi feita a remoção de atributos irrelevantes ou não pertinentes da base de dados, da discretização dos atributos numéricos em intervalos e transformação de outros em categorias, mantendo assim apenas atributos nominais, visando aumentar a possibilidade de algoritmos disponíveis na ferramenta Weka para a tarefa de classificação.

Para a mineração de processos, a própria ferramenta Disco possibilita excluir

da análise atributos indesejados ou até mesmo alterar o formato de atributos numéricos (como data e hora). Isso facilitou o processo de ajuste da base e fez com que a mineração de processo acontecesse de maneira fácil e rápida.

O quarto objetivo específico (realizar a análise descritiva da base de dados e analisar os resultados obtidos) foi alcançado através da análise da frequência dos valores para os atributos nominais utilizados na base. Verificou-se que alguns apresentam grande concentração em um grupo pequeno de valores e outros, uma distribuição mais uniforme. Esta análise da base de dados permitiu levar em conta essa concentração dos atributos ao avaliar os resultados obtidos na mineração de dados e mineração de processos.

5.2 TRABALHOS FUTUROS

Para trabalhos futuros, sugere-se validação dos resultados obtidos com usuários do sistema e especialistas da área de seguros. Sugere-se ainda a aplicação do estudo em outras bases de dados que envolvam fluxos de processos (sejam eles da área de seguros ou não) para possíveis comparações de resultados obtidos em bases com aspectos diferentes.

Além disso, é recomendada a aplicação de outras técnicas ou ferramentas de mineração de dados tradicional de forma a averiguar a aplicabilidade desta mineração em bases que envolvem processos e etapas interligadas.

REFERÊNCIAS

AMO, Sandra de. **Técnicas de mineração de dados**. Disponível em: <http://www.deamo.prof.ufu.br/arquivos/JAI-cap5.pdf>. Acesso em: 28 maio 2017.

BAEZA-YATES, Ricardo; RIBEIRO-NETO, Berthier. **Recuperação da informação:** conceitos e tecnologia das máquinas de busca. 2. ed. Porto Alegre: Bookman, 2013. 590 p.

CAMILO, Cássio Oliveira; SILVA, João Carlos da. **Mineração de dados:** conceitos, tarefas, métodos e ferramentas. Goiás: Instituto de Informática - Universidade Federal de Goiás, 2009. 29 p. (RT-INF_001-09). Disponível em: <http://www.inf.ufg.br/sites/*default*/files/uploads/relatorios-tecnicos/RT-INF_001-09.pdf>. Acesso em: 7 abr. 2017.

CASTRO, L. N. D.; FERRARI, D. G. **Introdução à mineração de dados**: conceitos básicos, algoritmos e aplicações. 1ª. ed. São Paulo: Saraiva, 2016.

FAYYAD, Usama et al. From *data mining* to knowledge discovery in databases. **AI magazine**, v. 17, n. 3, p. 37, 1996. Disponível em: <http://www.csd.uwo.ca/faculty/ling/cs435/fayyad.pdf>. Acesso em: 05 maio 2017.

FUNENSEG. Cadernos de Seguro. **Fundação Escola Nacional de Seguros**. Rio de Janeiro, v. 2, p. 27-29, 2001.

GALVÃO, N. D.; MARIN, H. D. F. Técnica de mineração de dados: uma revisão da literatura. **Acta Paul Enferm**, São Paulo, v. 22, n. 5, p. 686-690, 2009. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0103-21002009000500014>. Acesso em: 14 maio 2017.

KOHAVI, R. The Power of Decision Tables. **European Conference on Machine Learning**, vol. 912, p. 174-189. Disponível em: <https://pdfs.semanticscholar.org/ad13/187dc62e8dd39e767258c7e70767733d54e5.pdf>. Acesso em: 02 jun. 2017.

MEDEIROS, A. K. WEIJTERS, A. J. M. M. van der AALST, W.M.P. Genetic process mining: an experimental evaluation. **Data mining and knowledge discovery**, v.14, n.2, p. 245-304, Springer, 2007. Disponível em: <https://link.springer.com/article/10.1007/s10618-006-0061-7>. Acesso em: 25 abr. 2017.

POLETTO, G.A. **O Seguro Garantia em busca de sua natureza jurídica**. Rio de Janeiro: FUNENSEG, 2004.

SILVA, Edna Lúcia da; MENEZES, Estera Muszkat. **Metodologia da pesquisa e elaboração de dissertação.** 4. ed. Florianópolis: UFSC, 2005. 138 p. Disponível em: <https://projetos.inf.ufsc.br/arquivos/Metodologia_de_pesquisa_e_elaboracao_de_teses_e_dissertacoes_4ed.pdf>. Acesso em: 26 maio 2017.

van der AALST, W. M. P. WEIJTERS, A. Process Mining: a research agenda.

Computers in Industry. v. 53, n. 3, p. 231-244, 2004. Disponível em: <https://pure.tue.nl/ws/files/1965083/612473.pdf>. Acesso em: 25 abr. 2017.

van der AALST, Wil. **Process mining:** Data science in action. 2. ed. Springer, 2016.

WU, Xindong; KUMAR, Vipin. **The Top Ten Algorithms in Data Mining.** Chapman & Hall/CRC, 2009.